NIST Special Publication
NIST SP 1285

NIST Conference Papers
Fiscal Year 2020

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1285

NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication
NIST SP 1285

NIST Conference Papers
Fiscal Year 2020

Compiled and edited by:
Resources, Access, and Data Team
NIST Research Library and Museum

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1285

September 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

NIST SP 1285
September 2022

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

NIST Technical Series Policies

Publication History

Approved by the NIST Editorial Review Board on 2022-09-20

How to Cite this NIST Technical Series Publication

National Institute of Standards and Technology (2022) NIST Conference Papers Fiscal Year 2020. (U.S. Department of Commerce, Washington, D.C.), NIST Special Publication (SP) NIST SP 1285. https://doi.org/10.6028/NIST.SP.1285

Contact Information

techpubs@nist.gov .

Abstract

This Special Publication represents the work of researchers at professional conferences, as reported by NIST employees in Fiscal Year 2020 (October 1, 2019–September 30, 2020).

Keywords

NIST conference papers, NIST research, public access to NIST research.

Preface

NIST is committed to the idea that results of federally funded research are a valuable national resource and a strategic asset. To the extent feasible and consistent with law, agency mission, resource constraints, and U.S. national, homeland, and economic security, NIST will promote the deposit of scientific data arising from unclassified research and programs, funded wholly or in part by NIST, except for Standard Reference Data, free of charge in publicly accessible databases. Subject to the same conditions and constraints listed above, NIST also intends to make freely available to the public, in publicly accessible repositories, all peer-reviewed scholarly publications arising from unclassified research and programs funded wholly or in part by NIST.

This Special Publication represents the work of researchers at professional conferences, as reported in Fiscal Year 2020.

More information on public access to NIST research is available.

Table of Contents

Griffith, David W.; Ben Mosbah, Aziza. "On Upper Bounds for D2D Group Size." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 13, 2019. . . . . SP-1

Brubaker, Matthew; Spann, Bryan; Genter, Kristen; Roshko, Alexana; Blanchard, Paul; Harvey, Todd E.; Bertness, Kris A. "AlGaN/GaN core-shell heterostructures for nanowire UV LEDs." Paper presented at Nanowire Week, Hamilton, CA. June 11, 2018 - June 15, 2018. . . . . SP-7

Ng, Lisa; Kinser, Lindsey; Emmerich, Steven; Persily, Andrew K. "Estimating Interzonal Leakage in a Net-Zero Energy House." Paper presented at ASTM Symposium on Whole Building Air Leakage: Testing and Building Performance Impacts, San Diego, CA, US. April 08, 2018 - April 09, 2018. . . . . SP-8

Riddick, Frank H.; Wallace, Evan K.; Nieman, Scott; Tevis, Joe; Ferreyra, R. "Implementing Grain Traceability Standards: CART and Simulation." Paper presented at 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Detroit, MI, US. July 29, 2018 - August 01, 2018. . . . . SP-37

Kotsubo, Vincent; Ullom, Joel; Nam, Sae Woo. "Compact 1.7 K Cryocooler for Superconducting Nanowire Single-Photon Detectors." Paper presented at 20th International Cryocooler Conference, Burlington, VT, US. June 18, 2018 - June 21, 2018. . . . . SP-47

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018. . . . . SP-57

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018. . . . . SP-70

Sahoo, Anirudha; El Ouni, Naceur; Shenoy, Vineet. "A Study of Timing Constraints and SAS Overload in the CBRS Band using SAS-CBSD Protocol." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Wikola, HI, US. December 09, 2019 - December 13, 2019. . . . . SP-133

Shankar Raman, Arvind; Harper, Dustin; Haapala, Karl; Linke, Barbara; Bernstein, William Z.; Morris, KC. "Challenges In Representing

Manufacturing Processes for Systematic Sustainability Assessments - Workshop On June 21, 2018." Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, US. June 10, 2019 - June 14, 2019. . . . . SP-139

Sauca, Ana; Zhang, Chao; Seif, Mina; Choe, Lisa. "Evaluating Critical Temperatures of Axially Loaded I-shaped Steel Members Using ANSI/ AISC-360 Appendix 4 and Finite Element Model." Paper presented at Annual Stability Conference, St. Louis, MO, US. April 02, 2019 - April 05, 2019. . . . . SP-148

Rhorer, Richard; Quinn, George; Mates, Steven P. "Dynamic Compression Tests of Alumina Dumbbells Using a Spherical Joint." Paper presented at Society for Experimental Mechanics 2019 Annual Meeting, Reno, NV, US. June 03, 2019 - June 06, 2019. . . . . SP-161

Novotny, David. "Reducing Effects of LO Cable Movement in Antenna and Long Distance VNA Measurements." Paper presented at 2019 Joint International Symposium on Electromagnetic Compatibility, Sapporo and Asia-Pacific International Symposium on Electromagnetic Compatibility (EMC Sapporo/APEMC), Sapporo, JP. June 03, 2019 - June 07, 2019. . . . . SP-169

Mell, Peter; Delaitre, Aurelien; de Vaulx, Frederic; Dessauw, Philippe. " Implementing a Protocol Native Managed Cryptocurrency." Paper presented at The Fourteenth International Conference on Software Engineering Advances (ICSEA 2019), Valencia, GA, ES. November 24, 2019 - November 28, 2019. . . . . SP-173

Kelsey, John M.; Dachman-Soled, Dana; Sonmez Turan, Meltem; Mishra, Sweta. "TMPS: Ticket-Mediated Password Strengthening." Paper presented at The Cryptographer's Track of the RSA Conference (CT-RSA 2020), San Francisco, CA, US. February 24, 2020 - February 28, 2020. . . . . SP-181

Hassan, Md.; Chen, Wangchun. "Optimization of a Magnetostatic Cavity for a 3He Spin Analyzer on the CANDOR Polychromatic Reflectometer." Paper presented at 12th International Conference on Polarised Neutrons for Condensed Matter Investigations, Milton Hill, GB. July 03, 2018 - July 06, 2018. . . . . SP-204

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay. "On Practice and Theory of Constructive Composite Geometry and Topology." Paper presented at ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, US. August 18, 2019 - August 21, 2019. . . . . SP-215

Feng, Siyuan; Choi, Hyong-Ah; Griffith, David W.; Rouil, Richard A. "On

Selecting Channel Parameters for Public Safety Network Applications in LTE Direct." Paper presented at IEEE Consumer Communications & Networking Conference (CCNC), Las Vegas, NV. January 10, 2020 - January 13, 2020.. . . . . SP-227

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Throughput Analysis between Unicast and MBSFN from Link Level to System Level." Paper presented at 2019 IEEE 90th Vehicular Technology Conference: VTC2019-Fall, Honolulu, HI, US. September 23, 2019 - September 25, 2019.. . . . . SP-233

Wang, Jian; Rouil, Richard A.; Cintron, Fernando. "Distributed Resource Allocation Schemes for Out-of-Coverage D2D Communications." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 13, 2019.. . . . . SP-240

Hati, Archita; Nelson, Craig. "Multipath Mitigation from FM-AM Correlation." Paper presented at 2019 IEEE IFCS-EFTF, Orlando, FL, US. April 14, 2019 - April 18, 2019.. . . . . SP-250

Sheh, Raymond; Jacoff, Adam. "Best in Class: Leveraging Robot Performance Standards in Academic Competitions to Encourage Development and Dissemination." Paper presented at ASTM STP1614 on Symposium on Homeland Security and Public Safety: Research, Applications and Standards, San Diego, CA, US. June 28, 2018 - June 29, 2018.. . . . . SP-253

Leventon, Isaac; Bruns, Morgan. "Automated Fitting of Thermogravimetric Analysis Data." Paper presented at Interflam 2019, London, UK. July 01, 2019 - July 03, 2019.. . . . . SP-290

Leventon, Isaac; Bonny, Justin. "Influence of Dispositional and Situational Factors on Human Perceptions of Fire Risk." Paper presented at Interflam 2019, London, UK. July 01, 2019 - July 03, 2019.. . . . . SP-299

Sayrafian, Kamran; Krhac, Katjana; Noetscher, Gregory; Simunic, Dina. "A Simulation Platform to Study the Human Body Communication Channel." Paper presented at IEEE Engineering in Medicine and Biology Conference (EMBC 2019), Berlin, DE. July 23, 2019 - July 26, 2019.. . . . . SP-311

Tosado, Jacob; Chen, Wangchun; Rodriguez, Efrain E. "A Strategy for Handling Aberration in Spherical Neutron Polarimetry." Paper presented at 12th International Conference on Polarised Neutrons for Condensed Matter Investigations Journal of Physics: Conference Series, Abingdon, GB. July 03, 2018 - July 06, 2018. . . . . SP-315

Feng, Shaw C.; Lu, Yan; Jones, Albert T. "Data Registration for In-Situ Monitoring of Laser Powder Bed Fusion Processes." Paper presented at International Mechanical Engineering Congress and Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.. . . . . SP-325

Piliptchak, Pavlo; Aksu, Murat; Proctor, Frederick; Michaloski, John L. " Physics-based Simulation of Agile Robotic Systems." Paper presented at International Mechanical Engineering Congress & Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.. . . . . SP-335

Moges, Tesfaye; Witherell, Paul; Ameta, Gaurav. "On Characterizing Uncertainty Sources in Laser Powder Bed Fusion Additive Manufacturing Models." Paper presented at International Mechanical Engineering Congress and Exposition (IMECE2019), Salt Lake City, UT, US. November 11, 2019 - November 14, 2019. . . . . SP-343

Shao, Guodong; Jain, Sanjay; Laroque, Christoph; Lee, Loo Hay; Lendermann, Peter; Rose, Oliver. "Digital Twin for Smart Manufacturing: the Simulation Aspect." Paper presented at Winter Simulation Conference 2019, National Harbor, MD, US. December 08, 2019 - December 11, 2019.. . . . . SP-358

Tchoua, Roselyne; Ajith, Aswathy; Hong, Zhi; Ward, Logan; Chard, Kyle; Audus, Debra; Patel, Shrayesh; de Pablo, Juan. "Active Learning Yields Better Training Data for Scientific Named Entity Recognition." Paper presented at 2019 IEEE 15th International Conference on e-Science (e-Science), San Diego, CA, US. September 24, 2019 - September 27, 2019. . . . . SP-372

Shrestha, Som; Desjarlais, Andre; Dalgleish, Laverne; Ng, Lisa; Hun, Diana; Emmerich, Steven; Accawi, Gina. "Moisture Transfer in Commercial Buildings due to the Air Leakage: A New Feature in the Online Airtightness Savings Calculator." Paper presented at 2019 Buildings XIV International Conference, Clearwater, FL, US. December 09, 2019 - December 12, 2019.. . . . . SP-382

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun. " Infrastructure for Model Based Analytics for Manufacturing." Paper presented at 2019 Winter Simulation Conference, National Harbor, MD, US. December 08, 2019 - December 11, 2019.. . . . . SP-397

Souryal, Michael; Nguyen, Thao T. "Effect of Federal Incumbent Activity on CBRS Commercial Service." Paper presented at IEEE International Symposium on Dynamic Spectrum Access Networks, Newark, NJ, US. November 11, 2019 - November 14, 2019. . . . . SP-409

Vassilev, Apostol. "BowTie - a deep learning feedforward neural network for sentiment analysis." Paper presented at LOD 2019, Siena, IT. September 10, 2019 - September 13, 2019. . . . . SP-414

Schneider, Barry I.; Bartschat, Klaus; Zatsarinny, Oleg; Bray, Igor; Martin, Fernando; Scrinzi, Armin; Pamidighantam, Sudhakar; Tennyson, Jonathan; Gorfinkiel, Jimena; Klinker, Markus. "A Science Gateway for Atomic and Molecular Physics." Paper presented at Practice and Experience in Advanced Research Computing (PEARC), Chicago, IL, US. July 28, 2019 - August 01, 2019.. . . . . SP-426

Gao, Weichao; Sahoo, Anirudha. "Performance Study of a GAA-GAA Coexistence Scheme in the CBRS Band." Paper presented at IEEE Dynamic Spectrum Access Networks (DySPAN) Conference, Newark, NJ, US. November 11, 2019 - November 14, 2019.. . . . . SP-434

Potteiger, Bradley; Neema, Himanshu; Koutsoukos, Xenofon; Tang, CheeYee; Stouffer, Keith A. "Simulation Testbed for Railway Infrastructure Security and Resilience Evaluation." Paper presented at Hot Topics in the Science of Security (HotSoS) 2020, Lawrence, KS, US. April 07, 2020 - April 08, 2020. . . . . SP-444

Hall, Timothy; Caromi, Raied; Souryal, Michael; Wunderlich, Adam. " Reference Data Sets for Training and Evaluating RF Signal Detection and Classification Models." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 14, 2019.. . . . . SP-452

Jargon, Jeffrey; Long, Chris; Feldman, Ari; Martens, Jon. "Developing Models for a 0.8 mm Coaxial VNA Calibration Kit within the NIST Microwave Uncertainty Framework." Paper presented at 94th ARFTG Microwave Measurement Conference, San Antonio, TX, US. January 26, 2020 - January 29, 2020.. . . . . SP-457

Gamboa Quintiliani, Samantha; Moreaux, Alexandre; Griffith, David W.; Rouil, Richard A. "UE-to-Network Relay Discovery in ProSe-enabled LTE Networks." Paper presented at International Conference on Computing, Networking and Communications (ICNC 2020), Big Island, HI, US. February 17, 2020 - February 20, 2020.. . . . . SP-461

Phan, Long; Sadek, Fahim H.; Thonstad, Travis; Lew, Hai; Marcu, Sorin; Philip, Jacob. "Effects of Alkali-Silica Reaction on Mechanical Properties and Structural Capacities of Reinforced Concrete Structures." Paper presented at Structural Mechanics in Reactor Technology (SMiRT) 25, Charlotte, NC, US. August 04, 2019 - August 09, 2019.. . . . . SP-468

Quelennec, Aurore; Gorman, Jason J.; Reyes-Hernandez, Darwin. "Label-free surface acoustic wave-based embedded flow sensor." Paper presented at The 23rd International Conference on Miniaturized Systems for Chemistry and Life Sciences (µTAS 2019), Basel, CH. October 27, 2019 - October 31, 2019.. SP-477

Yuffa, Alex. "On Wacker•s Essential Equation in the Extrapolation Measurement Technique." Paper presented at Annual Symposium of the Antenna Measurement Techniques Association (AMTA), San Diego, CA, US. October 06, 2019 - October 11, 2019.. SP-479

Jargon, Jeffrey; Quimby, Jeanne; Remley, Kate; Koepke, Amanda; Williams, Dylan. "Verifying the Performance of a Correlation-Based Channel Sounder in the 3.5 GHz Band with a Calibrated Vector Network Analyzer." Paper presented at Asia-Pacific Microwave Conference (APMC) 2019, Singapore, SG. December 10, 2019 - December 13, 2019.. SP-484

Nguyen, Cuong; Roberts, Jeremy. "Green Button Data-Access Model for Smart Cities." Paper presented at The 2nd ACM/EIGSCC Symposium On Smart Cities and Communities, Portland, OR, US. September 10, 2019 - September 12, 2019.. SP-487

Sayrafian, Kamran; Roudneshin, Masoud; Aghdam, Amir. "Output Power Maximization of a Microgenerator Using Machine Learning Approach." Paper presented at The 7th Annual IEEE International Conference on Wireless for Space and Extreme Environments, Ottawa, CA. October 16, 2019 - October 18, 2019.. SP-493

Halba, Khalid; Griffor, Edward; Roth, Thomas; Kamongi, Patrick. "Using Statistical Methods and Co-Simulation to Evaluate ADS-Equipped Vehicle Trustworthiness." Paper presented at Electric Vehicles International Conference (EV) 2019, Bucharest, RO. October 03, 2019 - October 04, 2019.. SP-498

Persily, Andrew K.; Polidoro, Brian. "40th AIVC - 8th TightVent & 6th Venticool Conference, 2019 Residential Application of an Indoor Carbon Dioxide Metric." Paper presented at Conference 40th AIVC - 8th TightVent - 6th Venticool Conference in Ghent Belgium, Ghent, BE. October 15, 2019 - October 16, 2019.. SP-503

Mell, Peter. "Augmenting Fiat Currency with an Integrated Managed Cryptocurrency." Paper presented at The Fourteenth International Conference on Software Engineering Advances (ICSEA 2019), Valencia, ES. November 24, 2019 - November 28, 2019.. SP-516

Splett, Jolene D.; Jimenez, Felix; Koepke, Amanda. "Estimating the Parameters of Circles and Ellipses Using Orthogonal Distance Regression and Bayesian Errors-in-Variables Regression." Paper presented at 2019 Joint Statistical Meetings, Denver, CO, US. July 28, 2019 - August 01, 2019.. SP-524

Sieber, John. "How to Use and How Not to Use Certified Reference Materials in Industrial Chemical Metrology Laboratories." Paper presented at Denver X-Ray Conference, Lombard, IL, US. August 05, 2019 - August 09, 2019. SP-541

Persily, Andrew K. "40th AIVC - 8th Tight Vent & 6th Venticool Conference, 2019 Techniques to Estimate Commercial Building Infiltration Rates." Paper presented at 40th AIVC - 8th TightVent - 6th Venticool Conference, 2019, Ghent, BE. October 15, 2019 - October 16, 2019.. SP-558

Nablo, Brian; Ahn, Jungjoon; Bhadriraju, Kiran; Lee, Jong; Reyes-Hernandez, Darwin. "Live quantification of cell viability via neutral red uptake using lens-free imaging." Paper presented at The 23rd International Conference on Miniaturized Systems for Chemistry and Life Sciences (µTAS 2019), Basel, CH. October 27, 2019 - October 31, 2019. SP-567

Candell, Rick; Hany, Mohamed; Liu, Yongkang; Montgomery, Karl; Foufou, Sebti. "A Graph Database Approach to Wireless IIoT Work-cell Performance Evaluation." Paper presented at 2020 IEEE International Conference on Industrial Technology, Buenos Aires, AR. February 26, 2020 - February 28, 2020.. SP-569

Flowers-Jacobs, Nathan; Rufenacht, Alain; Fox, Anna; Waltman, Steven; Schwall, Robert; Brevik, Justus; Dresselhaus, Paul; Benz, Samuel P. " Development and Applications of a Four-Volt Josephson Arbitrary Waveform Synthesizer." Paper presented at 2019 IEEE International Superconductive Electronics Conference (ISEC), Riverside, CA, US. July 28, 2019 - August 01, 2019. SP-577

Fong, Jeffrey T.; Heckert, N. Alan; Filliben, James J.; Marcal, Pedro; Berweger, Samuel; Wallis, Thomas Mitchell (Mitch); Kabos, Pavel. " Design of an intelligent PYTHON code to run coupled and license-free finite-element and statistical analysis software for calibration of near-field scanning microwave microscopes." Paper presented at COMSOL Users Conference, Boston, MA, US. October 02, 2019 - October 04, 2019. SP-579

Wong, Vivian; Ferguson, Max; Law, Kincho; Lee, Yung-Tsun. "An Assistive Learning Workflow on Annotating Images for Object Detection." Paper presented at 2019 IEEE International Conference on Big Data, Los Angeles, CA, US. December 09, 2019 - December 12, 2019. SP-585

Falkenstein-Smith, Ryan; Sung, Kunhyuk; Chen, Jian; Hamins, Anthony. "

Chemical Structure of Medium-Scale Liquid Pool Fires." Paper presented at 13th International Symposium on Fire Safety Science, Waterloo, CA. April 27, 2020 - May 01, 2020.. . . . . SP-594

Persily, Andrew K.; Dols, William Stuart; Emmerich, Steven; Ng, Lisa. " The Role of Carbon Dioxide in Ventilation and IAQ Evaluation: 40 years of AIVC." Paper presented at 40th Air Infiltration and Ventilation Centre Conference, Ghent, BE. October 08, 2019 - October 09, 2019.. . . . . SP-610

Tam, Wai Cheong; Fu, Eugene Yujun; Mensch, Amy; Hamins, Anthony; You, Christina; Ngai, Grace; Leong, Hong va. "Prevention of Cooktop Ignition Using Detection and Multi-Step Machine Learning Algorithms." Paper presented at International Association of Fire Safety Science 2020, Waterloo, CA. April 27, 2020 - May 01, 2020.. . . . . SP-614

Huang, Siyuan; Hoskins, Brian; Daniels, Matthew; Stiles, Mark; Adam, Gina. "Streaming Batch Gradient Tracking for Neural Network Training." Paper presented at AAAI Conference on Artificial Intelligence, New York, NY, US. February 07, 2020 - February 12, 2020. . . . . . SP-637

Saidkhodjaev, Temur; Voas, Jeff; Kuhn, D. Richard; DeFranco, Joanna; Laplante, Phil. "Aggregating Atomic Clocks for Time-Stamping." Paper presented at 2020 IEEE International Conference on Service Oriented Systems Engineering (SOSE), Oxford, UK. August 03, 2020 - August 06, 2020.. . . . . SP-639

Koepke, Lucas; Splett, Jolene D.; Quinn, Timothy; Hrabe, Nik; Benzing, Jake; Frey, Michael. "Multiple Changepoint Analysis of Noisy Nonlinear Data with an Application to Modeling Crack Growth in Additively Manufactured Titanium." Paper presented at 2019 Joint Statistical Meetings, Denver, CO, US. July 28, 2019 - August 01, 2019.. . . . . SP-653

Sung, Kunhyuk; Chen, Jian; Bundy, Matthew; Hamins, Anthony. "The Characteristics of a 1 m Methanol Pool Fire." Paper presented at 13 International Symposium on Fire Safety Science (IAFSS 2020), Waterloo, CA. April 26, 2020 - May 01, 2020.. . . . . SP-664

Guo, Yang; Liu, Franklin; Wang, An; Liu, Hang. "AccuPIPE: Accurate Heavy Flow Detection in the Data Plane Using Programmable Switches." Paper presented at IEEE/IFIP Network Operations and Management, Budapest, HU. April 20, 2020 - April 24, 2020.. . . . . SP-682

McLinden, Mark O.; Kazakov, Andrei F.; Brown, J; Brignoli, Riccardo; Bell, Ian; Domanski, Piotr A. "Options for Low-GWP Refrigerants." Paper presented at Institute of Refrigeration, London, UK. November 06, 2019 - November 06, 2019.. . . . . SP-687

Zou, Qingtian; Singhal, Anoop; Sun, Xiaoyan; Liu, Peng. "Automatic Recognition of Advanced Persistent Threat Tactics for Enterprise Security." Paper presented at 6th ACM International Workshop on Security and Privacy Analytics 2020, New Orleans, LA, US. March 16, 2020 - March 18, 2020.. . . . . SP-705

Miller, Carl A. "The Impossibility of Efficient Quantum Weak Coin-Flipping." Paper presented at 52nd Annual ACM Symposium on Theory of Computing (STOC 2020), Chicago, IL, US. June 22, 2020 - June 26, 2020.. . . . . SP-717

Mosleh, Susanna; Ma, Yao; Rezac, Jake; Coder, Jason. "Dynamic Spectrum Access with Reinforcement Learning for Unlicensed Access in 5G and Beyond." Paper presented at 2020 IEEE 91st Vehicular Technology Conference: VTC2020-Spring, Antwerp, BE. May 25, 2020 - May 28, 2020. . . . . SP-741

Manzello, Samuel; Suzuki, Sayaka. "Influence of Angle Orientation on Firebrand Production from the Combustion of Materials Exposed to Applied Wind Fields." Paper presented at 57 Japanese Combustion Symposium, Hokkaido, JP. November 20, 2019 - November 22, 2019. . . . . SP-748

Manzello, Samuel; Suzuki, Sayaka. "Investigating coupled effect of radiative heat flux and firebrand showers on ignition of fuel beds." Paper presented at 57 Japanese Combustion Symposium, Sapporo (Hokkaido), JP. November 20, 2019 - November 22, 2019. . . . . SP-750

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda. "Forensic Analysis of Advanced Persistent Threat Attacks in Cloud Environments." Paper presented at Sixteenth IFIP 11.9 International Conference on Digital Forensics, New Delhi, IN. January 06, 2020 - January 08, 2020. . . . . SP-752

Park, Hyunseop; Ko, Hyunwoong; Lee, Yung-Tsun; Cho, Hyunbo; Witherell, Paul. "A Framework for Identifying and Prioritizing Data Analytics Opportunities in Additive Manufacturing." Paper presented at 5th International Workshop on Methodologies to Improve Managing Big Data projects in IEEE big data 2019, Los Angeles, CA, US. December 09, 2019 - December 12, 2019.. . . . . SP-773

Oe, Takehiko; Panna, Alireza; Elmquist, Randolph; Jarrett, Dean G.; Fukuyama, Yasuhiro; Kaneko, Nobu-Hisa. "Precise evaluation of the 129 kΩ and 1 MΩ quantum Hall array devices for a quantum Wheatstone bridge." Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.. . . . . SP-783

Ganesan, Adarsh; Seshia, Ashwin; Gorman, Jason J. "Phononic

Frequency Combs For Engineering MEMS/NEMS Devices With Tunable Sensitivity." Paper presented at IEEE SENSORS 2019, Montreal, CA. October 27, 2019 - October 30, 2019.. . . . . SP-785

Ma, Yao; Mosleh, Susanna; Coder, Jason. "A Unified Analytical Approach to Multi-Cell LBT-Based Spectrum Sharing Systems." Paper presented at the 2020 IEEE 91st Vehicular Technology Conference: VTC2020-Spring, Antwerp, BE. May 25, 2020 - May 28, 2020.. . . . . SP-789

Jun, Sung Yun; Caudill, Derek; Chuang, Jack; Papazian, Peter; Bodi, Anuraag; Gentile, Camillo; Senic, Jelena; Golmie, Nada T. "Penetration Loss at 60 GHz for Indoor-to-Indoor and Outdoor-to-Indoor Mobile Scenarios." Paper presented at IEEE European Conference on Antennas and Propagation, Copenhagen, DK. March 15, 2020 - March 20, 2020.. . . . . SP-795

Cohl, Howard; Greiner Petter, Andre; Schubotz, Moritz; Breitinger, Corinna; Muller, Fabian; Aizawa, Akiko; Gipp, Bela. "Discovering Mathematical Objects of Interest - A Study of Mathematical Notations." Paper presented at The Web Conference 2020, Taipei, TW. April 20, 2020 - April 24, 2020. . . . . SP-800

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Evaluating Unicast and MBSFN in Public Safety Networks." Paper presented at IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2020, London, UK. August 31, 2020 - September 03, 2020. . . . . SP-812

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.. . . . . SP-819

Mouha, Nicky; Celi, Christopher. "Extending NIST's CAVP Testing of Cryptographic Hash Function Implementations." Paper presented at CT-RSA 2020: Cryptographers' Track at the RSA Conference, San Francisco, CA, US. February 24, 2020 - February 28, 2020. . . . . SP-841

Gordon, Josh; Battou, Abdella; Kilper, Dan. "Workshop on Machine Learning for Optical Communication Systems: a summary." Paper presented at The Optical Networking and Communication Conference & Exhibition OFC 2020, San Diego, CA, US. March 08, 2020 - March 12, 2020.. . . . . SP-858

Copeland, Craig R.; McGray, Craig; Ilic, Robert; Geist, Jon; Stavis, Samuel M. "Particle Tracking of a Complex Microsystem in Three Dimensions and Six Degrees of Freedom." Paper presented at 2020 IEEE

33rd International Conference on Micro Electro Mechanical Systems (MEMS), Vancouver, CA. January 18, 2020 - January 22, 2020.. SP-861

Tait, Alexander; Buckley, Sonia; Shainline, Jeff; McCaughan, Adam; Chiles, Jeff; Nam, Sae Woo; Mirin, Richard. "Microresonator enhanced, waveguide coupled emission from silicon defect centers for superconducting optoelectronic networks." Paper presented at Optical Fiber Conference, San Diego, CA, US. March 08, 2020 - March 12, 2020. SP-865

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Cheung, Kin (Charles); Ye, Pei. "First Direct Measurement of Sub-Nanosecond Polarization Switching in Ferroelectric Hafnium Zirconium Oxide." Paper presented at IEEE International Electron Devices Meeting (IEDM), SAn Francisco, CA, US. December 07, 2019 - December 11, 2019. SP-874

Vernica, Teodor; Hanke, Aaron; Bernstein, William Z. "Leveraging standard geospatial representations for industrial augmented reality." Paper presented at MBE 2020: Model-Based Enterprise Summit 2020, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020. SP-878

Bhushan, Sumit; Slattery, Oliver T.; Tang, Xiao; Ma, Lijun. "Terahertz Electromagnetically Induced Transparency in Cesium Atoms." Paper presented at Frontier in Optics (FIO). September 14, 2020 - September 17, 2020. SP-885

Smith, Kassi; Bosworth, Bryan; Jungwirth, Nick; Cheron, Jerome; Orloff, Nate; Long, Chris; Williams, Dylan; Chamberlin, Richard; Quinlan, Franklyn; Fortier, Tara; Feldman, Ari. "On-Wafer Metrology of a Transmission Line Integrated Terahertz Source." Paper presented at CLEO: Laser Science to Photonic Applications, San Jose, CA, US. May 10, 2020 - May 15, 2020. SP-887

Xu, Hansong; Yu, Wei; Liu, Xing; Griffith, David W.; Golmie, Nada T. "On Data Integrity Attacks against Industrial Internet of Things." Paper presented at IEEE Cyber Science and Technology Congress (CyberSciTech 2020), Calgary, CA. August 17, 2020 - October 24, 2020.. SP-889

Wong, Vivian; Ferguson, Max; Law, Kincho; Lee, Yung-Tsun; Witherell, Paul. "Automatic Volumetric Segmentation of Additive Manufacturing Defects with 3D U-Net." Paper presented at SSS20 (VIRTUAL) AAAI Spring Symposium on AI in Manufacturing, Palo Alto, CA, US. March 23, 2020 - March 25, 2020.. SP-899

Diaz-Elsayed, Nancy; Hernandez, Luis; Rajamani, Ravi; Weiss, Brian A. " Asset Condition Management: A Framework for Smart, Health-Ready Manufacturing Systems." Paper presented at ASME 2020 15th

International Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020. . . . . SP-906

Weiss, Brian A.; Kaplan, Jared. "Assessment of a Novel Position Verification Sensor to Identify and Isolate Robot Workcell Health Degradation." Paper presented at ASME 2020 15th International Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020. . . . . SP-913

Dadfarnia, Mehdi; Sharp, Michael; Sprock, Timothy. "Understanding and Evaluating Naive Diagnostics Algorithms Applicable in Multistage Manufacturing from a Risk Management Perspective." Paper presented at ASME Manufacturing Science and Engineering Conference 2020 (MSEC 2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020. . . . . SP-922

Roth, Thomas; Nguyen, Cuong; Burns, Martin; Neema, Himanshu. " Integrating a Network Simulator with the High Level Architecture for the Co-Simulation of Cyber-Physical Systems." Paper presented at 2020 Simulation Innovation Workshop, Orlando, FL, US. February 10, 2020 - February 14, 2020.. . . . . SP-931

Wong, Kevin K F. "Identification of Geometric Nonlinearity Formulations and the Influences on Structural Dynamic Responses." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.. . . . . SP-943

Moody, Dustin; Perlner, Ray; Smith-Tone, Daniel; Apon, Daniel; Verbel, Javier. "Combinatorial Rank Attacks Against the Rectangular Simple Matrix Encryption Scheme." Paper presented at PQCrypto 2020: The Eleventh International Conference on Post-Quantum Cryptography, Paris, FR. April 15, 2020 - April 17, 2020.. . . . . SP-955

Hatada, Mitsuhiro; Scholl, Matthew. "Method for Effective Measurement, Labeling, and Classification of Botnet C2s for Predicting Attacks." Paper presented at 27th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, US. February 23, 2020 - February 26, 2020. . . . . SP-971

Dunkers, Joy; Da Silva, Sandra M.; Servetas, Stephanie; Filliben, James J.; Pinheiro, Guilherme; Lin, Nancy. "Key Considerations for Microbial Viability Measurements." Paper presented at Workshop on Ultraviolet Disinfection Technologies & Healthcare Associated Infections: Defining Standards and Metrology Needs, Gaithersburg, MD, US. January 14, 2020 - January 15, 2020.. . . . . SP-974

Kacker, Raghu N.; Kuhn, D. Richard. "CAgen: A fast combinatorial test generation tool with support for constraints and higher-index arrays." Paper presented at IEEE International Conference on Software Testing, Verification and Validation ICST 2020 Workshops, Porto, PT. March 23, 2020 - March 27, 2020.. SP-976

Sexton, Thurston; Brundage, Michael; Dima, Alden A.; Sharp, Michael. " MSEC: A Quantitative Retrospective." Paper presented at MSEC: Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020. SP-986

Amoah, Papa; Perez, Jesus; Obeng, Yaw S. "Application of Broadband RF Metrology to Integrated Circuit Interconnect Reliability Analyses: Monitoring Copper Interconnect Corrosion in 3D-ICs." Paper presented at 2020 International Conference on Microelectronic Test Structures, Edinburgh, UK. April 06, 2020 - April 09, 2020. SP-998

Nablo, Brian; Reyes-Hernandez, Darwin; Poster, Dianne L.; Postek, Michael; Obeng, Yaw S. "Standards, Metrology and Technology to Minimize Healthcare-Associated Infections: Novel approaches to measure efficacy." Paper presented at Workshop on Ultraviolet Disinfection Technologies & Healthcare Associated Infections: Defining Standards and Metrology Needs, Gaithersburg, MD, US. January 14, 2020 - January 15, 2020.. SP-1001

Haney, Julie; Furman, Susanne M.; Acar, Yasemin. "Smart Home Security and Privacy Mitigations: Consumer Perceptions, Practices, and Challenges." Paper presented at International Conference on Human-Computer Interaction, Copenhagen, DK. July 19, 2020 - July 24, 2020.. SP-1003

Wong, Kevin K F; McCabe, Steven. "Quantification of Seismic Energy Demand in Nonlinear Structures Through Analytical Derivations." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. SP-1022

Sadjadi, Omid; Greenberg, Craig; Singer, Elliot; Reynolds, Douglas; Mason, Lisa; Hernandez-Cordero, Jaime. "The 2019 NIST Speaker Recognition Evaluation CTS Challenge." Paper presented at The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, JP. November 01, 2020 - November 05, 2020. SP-1034

Artusio-Glimpse, Aly; Rogers, Kyle; Williams, Paul A.; Lehman, John H. " HALO - High Amplification Laser-pressure Optic." Paper presented at NEWRAD 2020, 14th International Conference on New Developments and Applications in Optical Radiometry, Boulder, CO, US. June 23, 2020 - June 26, 2020. SP-1041

Kant, Melis; Jaruga, Pawel; Coskun, Erdem; Ward, Samuel; Stark, Alexander; Becker, David; Adhikary, Amitav; Sevilla, Michael; Dizdar, Miral M. "Ion-beam radiation damage to DNA by investigation of free radical formation and base damage." Paper presented at DNA Damage, Mutation and Cancer; Gordon Research Conference, Ventura, CA, US. March 01, 2020 - March 06, 2020. . . . . SP-1043

Coskun, Erdem; Tuna, Gamze; Jaruga, Pawel; Tona, Alessandro; Erdem, Onur; Dizdar, Miral M. "Measurement of PARP1 in human tissues by liquid chromatography tandem mass spectrometry." Paper presented at DNA Damage, Mutation and Cancer, Gordon Research Conference, Ventura, CA, US. March 01, 2020 - March 06, 2020. . . . . SP-1044

Deisenroth, David; Mekhontsev, Sergey; Lane, Brandon. "Measurement of mass loss, absorbed energy, and time-resolved reflected power for laser powder bed fusion." Paper presented at SPIE Photonics West, San Francisco, CA, US. February 01, 2020 - February 06, 2020.. . . . . SP-1045

Cheung, Kin (Charles). "Anomalous accelerated negative-bias- instability (NBI) at low drain bias." Paper presented at 2020 International Reliability Physics Symposium, Virtual Dallas, TX, US. April 28, 2020 - May 30, 2020.. . . . . SP-1058

Sadjadi, Omid; Greenberg, Craig; Singer, Elliot; Reynolds, Douglas; Mason, Lisa; Hernandez-Cordero, Jaime. "The 2019 NIST Audio-Visual Speaker Recognition Evaluation." Paper presented at The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, JP. November 01, 2020 - November 05, 2020.. . . . . SP-1061

Kacker, Raghu N.; Kuhn, D. Richard. "Effectiveness of dataset reduction in testing machine learning algorithms." Paper presented at IEEE International Conference on Artificial Intelligence Testing, Oxford, UK. April 13, 2020 - April 16, 2020.. . . . . SP-1068

Elmquist, Randolph; Kruskopf, Mattias; Patel, Dinesh; Hu, I Fan; Liu, Chieh-I; Rigosi, Albert; Panna, Alireza; Payagala, Shamith; Jarrett, Dean G. "AC and DC Quantized Hall Array Resistance Standards." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1076

Segura, Christopher; Sattar, Siamak. "Uncertainty in the Seismic Response of Structural Components due to Material Variability." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. . . . . SP-1078

Georgakopoulos, Dimitrios; Budovsky, Ilya; Benz, Samuel P. "Josephson Arbitrary Waveform Synthesizer as a Reference Standard for the Calibration of Lock-in Amplifiers." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1090

Paul, Jagannath; Wahlstrand, Jared; Bristow, Alan. "Polarization-dependent absolute-phase-corrected multidimensional coherent spectra of exciton-polaritons." Paper presented at Ultrafast Phenomena and Nanophotonics XXIV, San Francisco, CA, US. February 03, 2020 - February 06, 2020. . . . . SP-1092

Mell, Peter; Gueye, Assane. "A Suite of Metrics for Calculating the Most Significant Security Relevant Software Flaw Types." Paper presented at 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, ES. July 13, 2020 - July 17, 2020. . . . . SP-1098

Fung, Juan; Sattar, Siamak; Butry, David; McCabe, Steven. "Machine Learning Methods for Predicting Seismic Retrofit Costs." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. . . . . SP-1104

Payagala, Shamith; Dee, Alana; Jarrett, Dean G. "Evaluation of an alternative null detector for adapted Wheatstone bridge." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 22, 2020 - August 28, 2020. . . . . SP-1114

Jun, Sung Yun; Caudill, Derek; Senic, Jelena; Gentile, Camillo; Chuang, Jack; Golmie, Nada T. "Over-The-Air Calibration of a Dual-Beam Dual-Polarized 28-GHz Phased-Array Channel Sounder." Paper presented at 2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting, Montreal, CA. July 05, 2020 - July 10, 2020. . . . . SP-1116

Oe, Takehiko; Payagala, Shamith; Jarrett, Dean G.; Kaneko, Nobu-Hisa. "Evaluation of NMIJ Traveling Dual Source Bridge Using NIST Adapted Wheatstone Bridge." Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1118

Yang, S.; Cular, Stefan; Rufenacht, Alain; Burroughs, Charles J.; Dresselhaus, Paul; Benz, Samuel P.; Ng, M. "Direct DC Voltages Comparison between two Programmable Josephson Voltage Standards at SCL." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1120

Bell, Ian. "Working Fluid Screening for Ocean Thermal Energy Conversion (OTEC) Applications." Paper presented at IIR International Rankine 2020 Conference - Heating, Cooling and Power Generation, Glasgow, UK, Glasgow, UK. July 26, 2020 - July 29, 2020. . . . . SP-1122

Bieler, Mark; Struszewski, Paul; Feldman, Ari; Jargon, Jeffrey; Hale, Paul D.; Gong, Pengwei; Xie, Wen; Yang, Chuntao; Feng, Zhigang; Zhao, Kejia; Yang, Zhijun. "Ultrafast waveform metrology: A first international comparison." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1128

Feldman, Ari; Jargon, Jeffrey; Dennis, Tasshi; Hale, Paul D. "Photodiode Calibration Comparison between Electro-Optic Sampling and Heterodyne Measurements up to 75 GHz." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1130

Kruskopf, Mattias; Patel, Dinesh; Liu, Chieh-I; Rigosi, Albert; Elmquist, Randolph; Wang, Yicheng; Bauer, Stefan; Yin, Yefei; Pierz, Klaus; Pesel, Eckard; Goetz, Martin; Schurr, Jurgen. "Graphene quantum Hall effect devices for AC and DC resistance metrology." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1132

Zhang, Shanshan; Lane, Brandon; Chou, Kevin. "Powder Thermal Conductivity Measurements in L-PBF using Powder-Included Build Specimens: Internal Geometry Effect." Paper presented at Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020. . . . . SP-1134

Rufenacht, Alain; Fox, Anna; Butler, Grace; Burroughs, Charles J.; Dresselhaus, Paul; Schwall, Robert; Cular, Stefan; Benz, Samuel P. " Compact DC Josephson Voltage Standard." Paper presented at 2020 Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1143

Geng, Jing; Li, Honglei; Hany, Mohamed; Liu, Yongkang; Candell, Rick; Bhattacharyya, Shuvra. "Integrating Field Measurements into a Model-Based Simulator for Industrial Communication Networks." Paper presented at 16th IEEE International Conference on Factory Communication Systems, Porto, PT. April 27, 2020 - April 29, 2020. . . . . SP-1145

Mates, Steven P.; Lopez-Hawa, Homar; Moscoso-Kingsley, Wilfredo;

Madhavan, Viswanathan. "Dynamic Flow Stress Measurements of 6061-T6 Aluminum under Rapid Heating for Machining Studies." Paper presented at Society for Experimental Mechanics 2020 Annual Meeting, Orlando, FL, US. June 08, 2020 - June 11, 2020.. . . . . . . SP-1153

Mosleh, Susanna; Ma, Yao; Rezac, Jake; Coder, Jason. "A Novel Machine Learning Approach to Estimating KPI and PoC for LTE-LAA-based Spectrum Sharing." Paper presented at IEEE International Conference on Communications (ICC) Workshop 2020, Dublin, IE. June 07, 2020 - June 11, 2020. . . . . . . SP-1159

Johnson, Ward L. "Contributions to Acoustic Loss in Langasite, Langatate, and Catangasite Resonators at High Temperatures." Paper presented at Sensor and Measurement Science International 2020, Nuremberg, DE. June 22, 2020 - June 25, 2020. . . . . . . SP-1165

Jarrett, Dean G.; Payagala, Shamith; Fitzgerald, Ryan P.; Bergeron, Denis E.; Cessna, Jeffrey T.; Waduwarage Perera, Charles; Zimmerman, Neil M. "Ohms Law Low-current Calibration System for Ionization Chambers." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . . . SP-1167

Gharavi, Hamid. "Deterministic Intra-Vehicle Communications: Timing and Synchronization." Paper presented at IEEE International Conference on Communications (IEEE ICC), Dublin, IE. June 07, 2020 - June 11, 2020.. . . . . . . SP-1169

Giraldo, Jairo; Tang, CheeYee; Urbina, David; Cardenas, Alvaro. "The More the Merrier: Adding Hidden Measurements for Anomaly Detection and Mitigation in Industrial Control Systems." Paper presented at Hot Topics in the Science of Security (HotSoS) 2020, Lawrence, KS, US. April 07, 2020 - April 08, 2020.. . . . . . . SP-1175

Fahnestock, Larry; Shi, Shitao; Speicher, Matthew. "Seismic Stability Assessment of Steel Moment Frames and Implications for Design." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. . . . . . . SP-1185

Maison, Bruce; Kasai, Kazuhiko; Speicher, Matthew. "Loading Protocols and Backbone Curves: U.S. and Japan Perspectives." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. . . . . . . SP-1197

Veksler, Dmitry; bersuker, gennadi; Bushmaker, A; Mason, Maribeth; Shrestha, Pragya; Cheung, Kin (Charles); Campbell, Jason; Rueckes, T;

Clevlend, L; Luan, H; Gilmer, D. "Memory update characteristics of carbon nanotube memristors (NRAM) under circuitry-relevant operation conditions." Paper presented at 2020 International Reliability Physics Symposium, Dallas, TX, US. April 28, 2020 - May 30, 2020.. . . . . .SP-1209

Lecci, Mattia; Testolina, Paolo; Giordani, Marco; Polese, Michele; Ropitault, Tanguy; Gentile, Camillo; Varshney, Neeraj; Zorzi, Michele. " Simplified Ray Tracing for the Millimeter Wave Channel: A Performance Evaluation." Paper presented at 2020 Information Theory and Applications Workshop, San Diego, CA, US. February 02, 2020 - February 07, 2020.. . . . . .SP-1213

Marangoni, Rafael; El Haddad, Darine; Seifert, Frank; Chao, Leon; Newell, David B.; Schlamminger, Stephan. "Measurement of the magnet system for the QEMMS." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.. . . . . .SP-1219

Marangoni, Rafael; El Haddad, Darine; Seifert, Frank; Chao, Leon; Pratt, Jon R.; Newell, David B.; Schlamminger, Stephan. "Design of the Kibble balance for the QEMMS." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.. . . . . SP-1221

Flowers-Jacobs, Nathan; Rufenacht, Alain; Fox, Anna; Dresselhaus, Paul; Benz, Samuel P. "Calibration of an AC Voltage Source Using a Josephson Arbitrary Waveform Synthesizer at 4 V." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020.. . . . . .SP-1223

Qiao, Guixiu; Garner, Jonathan. "Advanced sensing development to support accuracy assessment for industrial robot systems." Paper presented at ASME International Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020.. . . . . .SP-1225

Dukes, Jazalyn; Sattar, Siamak. "Development of a Database of Experimental Tests on FRP Retrofitted Reinforced Concrete Shear Walls." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. . . . . .SP-1232

Brevik, Justus; Soares Boaventura, Alirio De Jesus; Castellanos Beltran, Manuel; Donnelly, Christine; Flowers-Jacobs, Nathan; Fox, Anna; Hopkins, Pete; Dresselhaus, Paul; Williams, Dylan; Benz, Samuel P. "A Cryogenic Quantum-Based RF Source." Paper presented at 2020 95th

ARFTG Microwave Measurement Conference, Los Angeles, CA, US. August 04, 2020 - August 06, 2020.. . . . . SP-1241

Ayoade, Gbadebo; Akbar, Khandakar; Sahoo, Pracheta; Gao, Yang; Singhal, Anoop; Jee, Kangkook; Khan, Latifur; Agarwal, Anmol. " Evolving Advanced Persistent Threat Detection Using Provenance Graph and Metric Learning." Paper presented at IEEE International Conference on Communications and Network Security (CNS 2020), Avignon, FR. June 29, 2020 - July 01, 2020.. . . . . SP-1245

Sattar, Siamak; Johnson, Katherine. "Recommended Options for Improving the Functional Recovery of the Built Environment and Critical Infrastructure." Paper presented at 17 World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020. . . . . SP-1255

Flowers-Jacobs, Nathan; Babenko, Akim; Fox, Anna; Brevik, Justus; Dresselhaus, Paul; Benz, Samuel P. "Zero-Compensation Josephson Arbitrary Waveform Synthesizer at 1.33 V." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.. . . . . SP-1266

Manzello, Samuel. "Structure Ignition Mechanisms in Wildland-Urban Interface (WUI) Fire and Associated California State Fire Marshal Standard Test Methods." Paper presented at Japan Association for Fire Science and Engineering, Hirosaki, JP. May 28, 2020 - May 29, 2020.. . . . . SP-1268

Gopman, Daniel; Wang, Jian-Ping; Low, Tony; Wang, Weigang; Taylor, William. "Toward Ultra-fast and Ultra-low Power Switching in Perpendicular Magnetic Tunnel Junctions." Paper presented at GomacTech 2020 Conference, San Diego, CA, US. March 16, 2020 - March 20, 2020.. . . . . SP-1270

Hanke, Aaron; Vernica, Teodor; Bernstein, William Z. "Linking Performance Data and Geospatial Information of Manufacturing Assets through Standard Representations." Paper presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.. . . . . SP-1275

Caplins, Benjamin; Blanchard, Paul; Chiaramonti Debay, Ann; Diercks, David; Miaja Avila, Luis; Sanford, Norman A. "Correcting Systematic Energy Deficits in the Laser-Pulsed Atom Probe Mass Spectrum of SiO2." Paper presented at Microscopy and Microanalysis 2020, Milwaukee, WI, US. August 02, 2020 - August 06, 2020. . . . . SP-1284

Moges, Tesfaye; Yang, Zhuo; Jones, Kevontrez; Feng, Shaw C.;

Witherell, Paul; Lu, Yan. "Hybrid Modeling Approach for Melt Pool Prediction in Laser Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020. . . . . SP-1286

Marbukh, Vladimir; Sayrafian, Kamran; Rouzbehani, Behnam. "Utility Proportional Resource Allocation for Users with Diverse SLAs in Virtualized Radio Access Networks." Paper presented at IEEE International Black Sea Conference on Communications and Networking, Odesa, MD, US. May 26, 2020 - May 29, 2020. . . . . SP-1301

Feng, Shaw C.; Lu, Yan; Jones, Albert T. "Measured Data Alignments for Monitoring Metal Additive Manufacturing Processes Using Laser Powder Bed Fusion Methods." Paper presented at the ASME 2020 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, St. Louis, MO, US. August 16, 2020 - August 19, 2020. . . . . SP-1307

Remley, Kate; Bax, Clive; Mendivil, Edwin; Foegelle, Michael; Kvarnstrand, John; Skousen, Derek; Sanchez-Hernandez, David; Garcia-Fernandez, Miguel; Chang, Lorien; Yen, Elvis; Gutierrez, Johnny; Harbour, Justin. "Over-the-Air Testing of Cellular Large-Form-Factor Internet-of-Things Devices in Reverberation Chambers." Paper presented at 95th ARFTG Microwave Measurements Conference, Los Angeles, CA, US. June 26, 2020 - June 26, 2020. . . . . SP-1315

Manzello, Samuel; Suzuki, Sayaka. "Investigating Coupled Effect of Radiative Heat Flux and Firebrand Showers on Ignition of Fuel Beds - Structure vs Vegetative Firebrands." Paper presented at Japan Association for Fire Science and Engineering, Hirosaki, JP. May 28, 2020 - May 29, 2020. . . . . SP-1319

Curtis, Keith; Awad, George; Rajput, Shahzad; Soboroff, Ian. "HLVU: A new challenge to test deep understanding of movies the way humans do." Paper presented at ACM International Conference on Multimedia Retrieval (ICMR), Dublin, IE. June 08, 2020 - June 11, 2020. . . . . SP-1321

Landauer, Alexander; Van Blitterswyk, Jared; Riley, Michael A.; Forster, Aaron M. "Stress and Strain Heuristics for a Layered Elastomeric Foam at Medium Impact Rates." Paper presented at Society for Experimental Mechanics, Bethel, CT, US. September 14, 2020 - September 17, 2020. . . . . SP-1328

Barnes, Bryan; Henn, Mark-Alexander. "Contrasting Conventional and

Machine Learning Approaches to Optical Critical Dimension Measurements." Paper presented at Metrology, Inspection, and Process Control for Microlithography XXXIV, San Jose, CA, US. February 24, 2020 - February 27, 2020. . . . . SP-1333

Haney, Julie; Furman, Susanne M. "Towards Usable Updates for Smart Home Devices." Paper presented at 10th International Workshop on Socio-technical Aspects in Security, Guildford, UK. September 17, 2020 - September 17, 2020. . . . . SP-1346

Ibrahim, Kayla; Majumdar, Suryadipta; Bastos, Daniel; Singhal, Anoop. " Defining Actionable Rules for Verifying IOT Safety and Security." Paper presented at 41st IEEE Symposium on Security and Privacy 2020, San Francisco, CA, US. May 18, 2020 - May 20, 2020. . . . . SP-1361

Candell, Rick; Liu, Yongkang; Hany, Mohamed; Montgomery, Karl; Foufou, Sebti. "Smart Manufacturing Testbed for the Advancement of Wireless Adoption in the Factory." Paper presented at IFIP 17th International Conference on Product Lifecycle, Rapperswil, CH. July 05, 2020 - July 08, 2020. . . . . SP-1363

Lu, Xifeng; Gu, Dazhen; Kuester, Dan. "Digital Radiometer for Traceable Spectrum Sensing." Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020. . . . . SP-1377

Chao, Leon; Marangoni, Rafael; Seifert, Frank; El Haddad, Darine; Pratt, Jon R.; Newell, David B.; Schlamminger, Stephan. "The Design of an Instrument to Realize Small Torque at NIST." Paper presented at NCSLI (National Conference of Standards Laboratories International) Workshop & Symposium, Denver, CO, US. August 22, 2020 - August 27, 2020. . . . . SP-1379

Roth, Thomas; Lemieux, Christopher; Burns, Martin. "The UCEF Approach to Tool Integration for HLA Co-Simulations." Paper presented at Workshop on Design Automation for CPS and IoT, Sydney, AU. April 21, 2020 - April 21, 2020. . . . . SP-1385

Neema, Himanshu; Volgyesi, Peter; Koutsoukos, Xenofon; Roth, Thomas; Nguyen, Cuong. "Online Testbed for Evaluating Vulnerability of Deep Learning Based Power Grid Load Forecasters." Paper presented at Modeling and Simulation of Cyber-Physical Energy Systems, Sydney, AU. April 21, 2020 - April 21, 2020. . . . . SP-1394

Wang, Jun; Tam, Wai Cheong; Reneke, Paul A.; Peacock, Richard D.; Cleary, Thomas; Fu, Eugene Yujun; Ngai, Grace; Leong, Hong Va. "On the Use of Machine Learning Models to Forecast Flashover Occurrence in a Compartment." Paper presented at 17th International Conference on Automatic Fire Detection (AUBE 20) & Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2020), Mülheim an der Ruhr, DE. September 15, 2020 - September 17, 2020. . . . . SP-1400

Wang, Jun; Jia, Youwei; Fu, Eugene Yujun; Li, Jiajia; Tam, Wai Cheong. "Time Series Feature Extraction and Selection Tool for Fire Data." Paper presented at 17th International Conference on Automatic Fire Detection (AUBE 20) and Suppression, Detection and Signaling Research and Applications Conference (SUPDET 20), Mülheim an der Ruhr, DE. September 15, 2020 - September 17, 2020.. . . . . SP-1408

Falkenstein-Smith, Ryan; Cleary, Thomas. "Assessing Fire Smoke to Predict Backdraft and Smoke Explosion Potential." Paper presented at 17th International Conference on Automatic Fire Detection (AUBE 20) & Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2020), Mülheim an der Ruhr, DE. September 15, 2020 - September 17, 2020. . . . . SP-1416

Apon, Daniel; Perlner, Ray; Robinson, Angela; Santini, Paulo. " Cryptanalysis of LEDAcrypt." Paper presented at Crypto 2020, Santa Barbara, CA, US. August 16, 2020 - August 20, 2020. . . . . SP-1425

Barve, Yogesh; Neema, Himanshu; Kang, Zhuangwei; Sun, Hongyang; Gokhale, Aniruddha; Roth, Thomas. "EXPPO: EXecution Performance Profiling and Optimization for CPS Co-simulation-as-a-Service." Paper presented at the 23rd International Symposium on Real-Time Distributed Computing, Nashville, TN, US. May 19, 2020 - May 21, 2020.. . . . . SP-1455

Yang, Zhuo; Lu, Yan; Yeung, Ho; Krishnamurty, Sundar. "3D Build Melt Pool Predictive Modeling for Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020. . . . . SP-1463

Brundage, Michael; Pavel, Radu. "Using Text Analytics Solutions with Small to Medium Sized Manufacturers: Lessons Learned." Paper presented at Model-Based Enterprise Summit, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020. . . . . SP-1477

Brundage, Michael; Chandrasegaran, Senthil; Zhang, Xiaoyu; Ma, Kwan-Liu. "Using Text Visualization to aid Analysis of Machine Maintenance Logs." Paper presented at Model-Based Enterprise Summit, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020. . . . . SP-1486

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Campbell, Jason; Cheung, Kin (Charles); Ye, Peide. "Record Fast Polarization Switching Observed in Ferroelectric Hafnium Oxide Crossbar Arrays." Paper presented at IEEE Silicon Nanoelectronics Workshop, Honolulu, HI, US. June 13, 2020 - June 14, 2020.. . . . . SP-1497

Attota, Ravikiran. "Computational Process Control Compatible Dimensional Metrology Tool: Through-focus Scanning Optical Microscopy." Paper presented at 2020 31st Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), Saratoga Springs, NY, US. August 24, 2020 - August 26, 2020. . . . . SP-1499

Zwolak, Justyna; Taylor, Jacob; Kalantre, Sandesh; McJunkin, Thomas; Weber, Brian. "Ray-based classification framework for high-dimensional data." Paper presented at Thirty-fourth Conference on Neural Information Processing Systems, Vancouver, CA. December 05, 2020 - December 12, 2020. . . . . SP-1502

Kolmakov, Andrei; Tselev, Alexander. "Operando Scanning Electron and Microwave Microscopies in Plasmas: A Comparative Analysis." Paper presented at Microscopy and Microanalysis 2020, Milwaukee, WI, US. August 02, 2020 - August 06, 2020.. . . . . SP-1509

Forbes, Thomas P.; Krauss, Shannon; Jobes, Dillion. "Gradient Elution Moving Boundary Electrophoresis of Homemade Fuel-Oxidizer Explosives." Paper presented at MicroTAS 2020. October 04, 2020 - October 09, 2020.. . . . . SP-1511

Shi, Junxiao; Pesavento, Davide; Benmohamed, Lotfi. "NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware." Paper presented at 7th ACM Conference on Information-Centric Networking (ICN 2020), Montreal, CA. September 29, 2020 - October 01, 2020.. . . . . SP-1513

Lubell, Joshua. "A Document-based View of the Risk Management Framework." Paper presented at Balisage: The Markup Conference, Washington, DC, US. July 27, 2020 - July 31, 2020. . . . . SP-1524

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Quality of Service Optimization in Mobile Edge Computing Networks via Deep Reinforcement Learning." Paper presented at The 15th International Conference on Wireless Algorithms, Systems, and Applications (WASA 2020), Qingdao, CN. September 13, 2020 - September 15, 2020. . . . . SP-1544

Piez, Wendell. "Systems security assurance as (micro) publishing: Declarative markup for systems description and assessment." Paper presented at Balisage: The Markup Conference 2020, Rockville, MD, US. July 27, 2020 - July 31, 2020. . . . SP-1556

Kuo, Paulina. "Effect of Amplitude Mismatch on Entanglement Visibility in Photon-Pair Sources." Paper presented at Integrated Photonics Research, Silicon and Nanophotonics (IPR), Montreal, CA. July 13, 2020 - July 16, 2020. . . . SP-1573

Ivezic, Nenad; Jelisic, Elena; Kulvatunyou, Boonserm; Nieman, Scott; Oh, Hakju; Marjanovic, Zoran. "Towards inter-operable enterprise systems • graph-based validation of a context-driven approach for message profiling." Paper presented at APMS 2020 International Conference Advances in Production Management Systems (APMS 2020), Novi Sad, RS. August 30, 2020 - September 03, 2020. . . . SP-1576

Cho, Seulki; Le, Son; Richter, Curt A.; Balijepalli, Arvind. "High-Resolution Biochemical Activity Measurements with Commercial Transistors." Paper presented at MicroTAS 2020. October 04, 2020 - October 09, 2020. . . . SP-1584

Haney, Julie; Acar, Yasemin; Furman, Susanne M. "It's the Company, the Government, You and I: User Perceptions of Responsibility for Smart Home Privacy and Security." Paper presented at USENIX Security 2021, Vancouver, CA. August 11, 2021 - August 13, 2021. . . . SP-1586

Deisenroth, David; Hanssen, Leonard; Mekhontsev, Sergey. "High temperature reflectometer for spatially resolved spectral directional emissivity of laser powder bed fusion processes." Paper presented at Reflection, Scattering, and Diffraction from Surfaces VII, San Diego, CA, US. August 24, 2020 - August 28, 2020. . . . SP-1608

Gerrits, Thomas; Migdall, Alan; Bienfang, Joshua; Lehman, John H.; Nam, Sae Woo; Slattery, Oliver T.; Splett, Jolene D.; Vayshenker, Igor; Wang, Chih-Ming. "Calibration of free-space and fiber-coupled single-photon detectors." Paper presented at Quantum 2.0 Conference. September 14, 2020 - September 17, 2020. . . . SP-1628

Qu, Yongzhi; Vogl, Gregory W.; Wang, Zechao. "A Deep Neural Network Model for Learning Runtime Frequency Response Function Using Sensor Measurements." Paper presented at ASME 2021 16th International Manufacturing Science and Engineering Conference (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021. . . . SP-1630

NIST SP 1285
September 2022

On Upper Bounds for D2D Group Size

David Griffith and Aziza Ben Mosbah
National Institute of Standards & Technology
Gaithersburg, MD, USA
david.griffith@nist.gov

***Abstract*—In this paper, we derive upper bounds for the number of Device-to-Device (D2D)-capable out-of-coverage (OOC) User Equipments (UEs) that can share the Physical Sidelink Discovery Channel (PSDCH) while maintaining a minimum probability of discovery message decoding. We maximize these upper bounds with respect to the UEs' transmission probability threshold by exploiting the fact that the upper bound is nearly linear with respect to the number of resources in the discovery resource pool. The resulting simple approximate bound is accurate over a large range of parameter values. We validate our results using Monte Carlo simulations in MATLAB and the ns-3 simulation tool.**

I. Introduction

Device-to-device (D2D) communications offer a means for improving cellular network efficiency by reducing the load at the base station due to intra-cell traffic. The Proximity Services (ProSe) working group in the Third Generation Partnership Project (3GPP) has defined standards for D2D communications for UEs that are within a base station's coverage area and also for out-of-coverage (OOC) UEs, i.e., those that are outside any cellular coverage [1]. The latter case affects public safety users who may be deployed to remote areas or who may have to operate in regions where cellular service is offline due to a natural or other causes.

Each User Equipment (UE) in a group of UEs must discover the D2D applications hosted by other UEs in the group before it can establish D2D sessions with them. OOC UEs implement the discovery function by transmitting discovery messages over the Physical Sidelink Discovery Channel (PSDCH). The UEs randomly choose Physical Resource Block (PRB) pairs from a periodically recurring discovery resource pool whose dimensions are N_t, the number of subframe sets allocated for a transmission, by N_f, the number of PRB pairs in the frequency domain. There are $N_r = N_f N_t$ PRB pairs, i.e., PSDCH resources, in the discovery resource pool. Because they choose resources randomly, multiple UEs can select the same resource, which causes message collisions that reduce the PSDCH's throughput and delay UE discovery. The ProSe standard tries to fix this problem by incorporating a transmission probability threshold, θ, that UEs can use to throttle their transmissions; a UE will transmit on the PSDCH only if an internally randomly generated number in the unit interval $[0,1]$ is less than θ [2, Clause 5.15.1.1]. The value of θ determines how many UEs can participate in the D2D discovery process while maintaining a given level of performance, i.e., at least a minimum discovery message decoding probability.

Our previous work found the optimal value of θ that maximizes the probability of a successful discovery message transmission from one UE to another [3]. In this paper, we extend this work by deriving an expression for the maximum number of OOC UEs whose members' probability of decoding discovery messages is above a given minimum threshold. In Section II, we obtain an expression for the group size, N_u, as a function of other PSDCH parameters and use it to derive the maximum group size that supports at least a minimum discovery message decoding probability, $P_{\min}$. In Section III, we maximize the group size upper bound with respect to θ. Because direct methods are not tractable, we obtain approximate closed-form expressions for the critical value of θ, and use them to get the maximum group size upper bound. In Section IV, we validate our model using simulations and discuss the implications of the results, and in Section V, we summarize our results.

II. The Maximum Group Size

Let N_u be the number of OOC UEs. We assume that every UE can communicate with every other UE in its neighborhood, and that the UEs are using Model A discovery, i.e., they are continuously sending discovery announcement messages. In each discovery period, each UE generates a discovery message, selects a pool resource, and transmits its message with probability θ. We randomly select two UEs from their group and designate them as UE X and UE Y. From [3], $P_{Y\to X}$, the probability that UE X successfully decodes UE Y's discovery message during a given period, is

$$P_{Y\to X} = \theta\left(1-\frac{\theta}{N_t}\right)\left(1-\frac{\theta}{N_r}\right)^{N_u-2}. \qquad (1)$$

Solving Eq. (1) for N_u gives

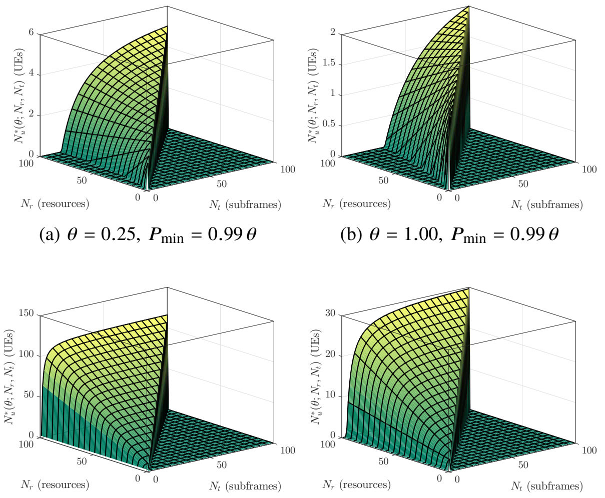
$$N_u = \left\lfloor 2+\frac{\log\left(P_{Y\to X}/\theta\right)-\log\left(1-\theta/N_t\right)}{\log\left(1-\theta/N_r\right)}\right\rfloor, \qquad (2)$$

Since $P_{Y\to X} \leq 1$, $\theta \leq 1$, and both N_t and N_r are positive integers, it follows from Eq. (1) that $P_{Y\to X} < \theta$ and $\log\left(P_{Y\to X}/\theta\right) < 0$.

Let $P_{\min}$ be the minimum allowable value of $P_{Y\to X}$. We are interested in determining the largest number of UEs that the

Disclaimer: Certain commercial products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the commercial products identified are necessarily the best available for the purpose.

Griffith, David W.; Ben Mosbah, Aziza. "On Upper Bounds for D2D Group Size." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 13, 2019.



(a) $\theta = 0.25,\ P_{\min} = 0.99\,\theta$

(b) $\theta = 1.00,\ P_{\min} = 0.99\,\theta$

(c) $\theta = 0.25,\ P_{\min} = 0.75\,\theta$

(d) $\theta = 1.00,\ P_{\min} = 0.75\,\theta$

Fig. 1. Plots of $N_u^*(\theta; N_r, N_t)$ versus N_t and N_r, for various values of θ and $P_{\min}$.

discovery resource pool can support so that $P_{Y\to X} \ge P_{\min}$. The derivative of the argument of the floor function in Eq. (2) with respect to $P_{Y\to X}$ is $\left(P_{Y\to X}\log(1-\theta/N_r)\right)^{-1}$. Since $\log(1-\theta/N_r) < 0$, the argument of the floor function is monotonically decreasing with respect to $P_{Y\to X}$. Therefore the maximum value of N_u given $P_{Y\to X} \ge P_{\min}$ is

$$N_u^*(\theta; N_r, N_t) = \left\lfloor 2 + \frac{\log\left(P_{\min}/\theta\right) - \log\left(1-\theta/N_t\right)}{\log\left(1-\theta/N_r\right)} \right\rfloor. \quad (3)$$

Eq. (3) does not hold if $N_t > N_r$, since it is not possible for the number of subframes spanned by the resource pool to exceed the number of pool resources. Also, if $N_u^*(\theta; N_r, N_t) < 0$, then $P_{Y\to X} < P_{\min}$ for all values of N_u. This occurs when

$$N_t < \frac{\theta^2}{\theta - P_{\min}(1-\theta/N_r)^2} \le \frac{\theta^2}{\theta - P_{\min}}, \quad (4)$$

where the higher upper bound is the limit of the lower upper bound as $N_r \longrightarrow \infty$. Eq. (4) indicates that $P_{\min} > P_{Y\to X}$ for all values of N_u when N_t is sufficiently small, because of the half duplex effect.

Fig. 1 shows the effect of variations in θ and $P_{\min}$ on $N_u^*(\theta; N_r, N_t)$. In every subfigure in Fig. 1, we set $N_u^*(\theta; N_r, N_t) = 0$ when $N_t > N_r$, because the number of resources in the pool cannot be less than the number of subframes spanned by the pool. Also, when Eq. (4) holds, we set $N_u^*(\theta; N_r, N_t) = 0$, which produces the flat regions on the left side of each plot in Fig. 1a and Fig. 1b. Also, setting $P_{\min}$ close to θ reduces $N_u^*(\theta; N_r, N_t)$; for example, Fig. 1b shows that when $N_r = 100$ resources, at most two UEs can achieve $P_{Y\to X} \ge 0.99$ when $\theta = 1$.

Fig. 1 shows that the slope of the $N_u^*(\theta; N_r, N_t)$ surface is nearly linear in the N_r direction when N_r and N_t are large. The first and second order partial derivatives of $N_u^*(\theta; N_r, N_t)$ with respect to N_r are

$$\frac{\partial N_u^*(\theta; N_r, N_t)}{\partial N_r} = -\frac{\theta\big(\log(P_{\min}/\theta) - \log(1-\theta/N_t)\big)}{N_r(N_r-\theta)\log^2(1-\theta/N_r)} \quad (5)$$

and

$$\frac{\partial^2 N_u^*(\theta; N_r, N_t)}{\partial N_r^2} = \big(\log(P_{\min}/\theta) - \log(1-\theta/N_t)\big) \times \frac{2\theta^2 + (2\theta N_r - \theta^2)\log(1-\theta/N_r)}{N_r^2(N_r-\theta)^2\log^3(1-\theta/N_r)}. \quad (6)$$

From Eq. (5), when N_r is large the derivative approaches

$$\lim_{N_r\to\infty} \frac{\partial N_u^*(\theta; N_r, N_t)}{\partial N_r} = \frac{\log(1-\theta/N_t) - \log(P_{\min}/\theta)}{\theta}. \quad (7)$$

As N_t increases, the slope in Eq. (7) approaches a constant value: $\lim_{N_r\to\infty, N_t\to\infty} \partial N_u^*(\theta; N_r, N_t)/\partial N_r \approx -\log(P_{\min}/\theta)/\theta$. We confirm this by letting $N_r \longrightarrow \infty$ in Eq. (6), which gives $\lim_{N_r\to\infty} \partial^2 N_u^*(\theta; N_r, N_t)/(\partial N_r^2) = 0$.

Thus, for any value of N_t, we can approximate $N_u^*(\theta; N_r, N_t)$ as a linear function of N_r. We use a Taylor series expansion about the point $N_r = N_t$. If we retain only the constant and linear terms, we get

$$N_u^*(\theta; N_r, N_t) \approx \left(1 + \frac{\log(P_{\min}/\theta)}{\log(1-\theta/N_t)}\right) - \frac{\theta\left(\log(P_{\min}/\theta) - \log(1-\theta/N_t)\right)}{N_t(N_t-\theta)\log^2(1-\theta/N_t)}(N_r - N_t). \quad (8)$$

We want to find the maximum group size, given a pool of N_r resources, over the range of values for N_t. We define this upper bound as follows:

$$N_u^{\max}(\theta; N_r) = \max_{N_t} N_u^*(\theta; N_r, N_t). \quad (9)$$

We can use Eq. (3) to derive $N_u^{\max}(\theta; N_r)$. The derivative of the argument of the floor function in Eq. (3) with respect to N_t is $-\theta/[N_t(N_t-\theta)\log(1-\theta/N_r)]$. Since $\log(1-\theta/N_r) < 0$ and $\theta < N_t \le N_r$, the derivative is positive over the interval of interest, so that $N_u^*(\theta; N_r, N_t)$ is a non-decreasing function of N_t. Setting $N_t = N_r$ gives

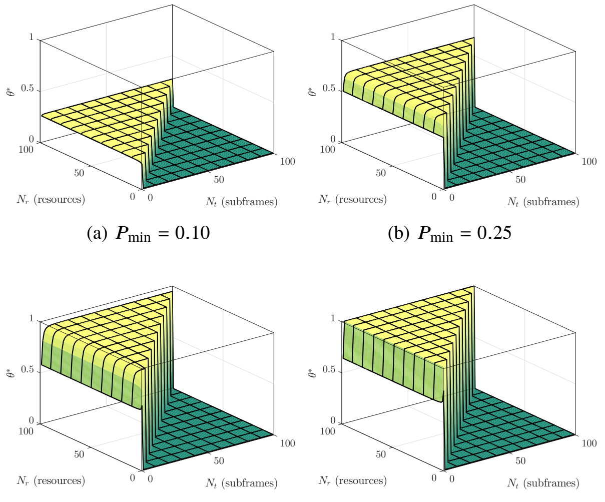
$$N_u^{\max}(\theta; N_r) = \left\lfloor 1 + \frac{\log\left(P_{\min}/\theta\right)}{\log\left(1-\theta/N_r\right)} \right\rfloor \ge N_u^*(\theta; N_r, N_t). \quad (10)$$

III. Maximum Group Size Bounds with Respect to θ

In this section, we maximize N_u^* in Eq. (3) and $N_u^{\max}$ in Eq. (10) with respect to θ. We assume that all N_u UEs use the same value for θ. We let θ^* be the value of θ that maximizes $N_u^*(\theta; N_r, N_t)$, and we let $\theta^{\max}$ be the value of θ that maximizes $N_u^{\max}(\theta; N_r)$.

We can simplify the problem by showing that θ^* is nearly constant with respect to N_r and N_t, and by showing that $\theta^* \approx \theta^{\max}$ unless N_t is small. Consider a set of pool dimensions where $N_r \le 100$ resources and $N_t \le N_r$ subframes. For each ordered pair (N_r, N_t), we use numerical methods to find θ^*,

Griffith, David W.; Ben Mosbah, Aziza. "On Upper Bounds for D2D Group Size." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 13, 2019.



(a) $P_{\min} = 0.10$ (b) $P_{\min} = 0.25$

(c) $P_{\min} = 0.35$ (d) $P_{\min} = 0.50$

Fig. 2. Plots of θ^*, which is the value of θ that minimizes Eq. (3), versus N_t and N_r, for various values of $P_{\min}$.

and we use the following values of $P_{\min}$: 0.10, 0.25, 0.35, and 0.50; the resulting plots of θ^* are shown in Fig. 2.

Fig. 2 shows that θ^* is insensitive to N_r and N_t, except when N_t is small (i.e., $N_t < 10$ subframes). Fig. 2a shows that when $P_{\min}$ is small, θ^* is nearly constant over all values of N_r and N_t, and Fig. 2d shows that θ^* varies with respect to N_t only when $N_t = 1$ subframe. Fig. 2b and Fig. 2c show more variation of θ^* with respect to N_t, but this occurs when $N_t < 10$ subframes.

The insensitivity of θ^* to N_r and N_t, and the fact that $\theta^* = \theta^{\max}$ when $N_r = N_t$, implies that $\theta^* \approx \theta^{\max}$ unless N_t is very small. Additional analysis shows that if $P_{\min} = 1/\mathsf{e}$, $\theta^* = 1$ over almost the entire set of values for N_t and N_r. Thus, we can obtain a good approximation for θ^* by finding $\theta^{\max}$.

As we will show, $N_u^{\max}(\theta; N_r)$ is linear with respect to the pool size, N_r. We expand $N_u^{\max}(\theta; N_r)$ in a Taylor series about $N_r = a$:

$$\begin{aligned} N_u^{\max}(\theta; N_r) = \Big\lfloor \Big(1 + \tfrac{\log(P_{\min}/\theta)}{\log(1-\theta/a))}\Big) - \tfrac{\theta \log(P_{\min}/\theta)}{a(a-\theta)\log^2(1-\theta/a))}(N_r - a) \\ + \tfrac{1}{2}\tfrac{\log(P_{\min}/\theta)\left[2\theta^2-(\theta^2-2a\theta)\log(1-\theta/a)\right]}{a^2(a-\theta)^2\log^3(1-\theta/a)}(N_r - a)^2 \\ + O\big((N_r - a)^3\big)\Big\rfloor. \end{aligned} \quad (11)$$

Derivatives of order two and higher in Eq. (11) vanish as a becomes very large, so the limit as $a \to \infty$ is

$$\begin{aligned} \lim_{a\to\infty} N_u^{\max}(\theta; N_r) &= \Big\lfloor \lim_{a\to\infty}\Big(1 + \tfrac{[\theta+(a-\theta)\log(1-\theta/a)]\log(P_{\min}/\theta)}{(a-\theta)\log^2(1-\theta/a)}\Big) \\ &\qquad - N_r \lim_{a\to\infty} \tfrac{\theta\log(P_{\min}/\theta)}{a(a-\theta)\log^2(1-\theta/a)}\Big\rfloor \\ &= \Big\lfloor 1 + \Big(\tfrac{1}{2} - \tfrac{N_r}{\theta}\Big)\log(P_{\min}/\theta)\Big\rfloor. \end{aligned} \quad (12)$$

To find our approximate value for $\theta^{\max}$, we take the derivative of the argument of the floor function with respect to θ:

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\Big[1 + \Big(\tfrac{1}{2} - \tfrac{N_r}{\theta}\Big)\log\big(\tfrac{P_{\min}}{\theta}\big)\Big] = \frac{2N_r(1 + \log(P_{\min}/\theta)) - \theta}{2\theta^2}. \quad (13)$$

Setting the numerator in Eq. (13) equal to zero and rearranging the resulting expression gives:

$$\frac{\mathsf{e}P_{\min}}{2N_r} = \frac{\theta^{\max}}{2N_r}\mathsf{e}^{\theta^{\max}/(2N_r)}. \quad (14)$$

To solve Eq. (14) for $\theta^{\max}$, we use Lambert's W function, $\mathsf{W}(z)$, which is defined $\forall z \in \mathbb{C}$ as $z = \mathsf{W}(z)\exp\big(\mathsf{W}(z)\big)$ [4, Eq. (1.5)]. We match Eq. (14) to the W function's definition by letting $z = \mathsf{e}P_{\min}/(2N_r)$ and $\mathsf{W}(z) = \theta^{\max}/(2N_r)$, giving

$$\theta^{\max} = 2N_r\,\mathsf{W}(z) = 2N_r\,\mathsf{W}\big(\mathsf{e}P_{\min}/(2N_r)\big). \quad (15)$$

We also obtain a simple approximate value for $\theta^{\max}$ as follows. The Taylor series expansion of the W function about the point $x = 0$ has the form [4, Eq. (3.1)]

$$\mathsf{W}(x) = \sum_{n=1}^{\infty}\frac{(-n)^{n-1}}{n!}x^n, \quad (16)$$

and from this and Eq. (15), we get the series expansion for $\theta^{\max}$:

$$\theta^{\max} = \mathsf{e}P_{\min} - \frac{(\mathsf{e}P_{\min})^2}{2N_r} + O\left(\frac{(P_{\min})^3}{N_r^2}\right). \quad (17)$$

Since $P_{\min} > 0$ and $\mathsf{W}(0) = 0$, letting $N_r \to \infty$ gives

$$\lim_{N_r\to\infty}\theta^{\max} = \mathsf{e}P_{\min}\exp\big(-\mathsf{W}(0)\big) = \mathsf{e}P_{\min}, \quad (18)$$

Thus we have the following approximation for $\theta^{\max}$:

$$\theta^{\max} \approx \begin{cases} \mathsf{e}P_{\min}, & 0 < P_{\min} \le 1/\mathsf{e} \\ 1, & 1/\mathsf{e} < P_{\min} \le 1 \end{cases} \quad (19)$$

Using Eq. (19) with Eq. (10), we get the following approximate expression for the maximum group size:

$$N_u^{\max}(\theta; N_r) \approx \begin{cases} \left\lfloor 1 - \frac{1}{\log\left(1 - \frac{\mathsf{e}P_{\min}}{N_r}\right)}\right\rfloor, & 0 < P_{\min} \le 1/\mathsf{e} \\ \left\lfloor 1 + \frac{\log(P_{\min})}{\log\left(1 - \frac{1}{N_r}\right)}\right\rfloor, & 1/\mathsf{e} < P_{\min} \le 1 \end{cases} \quad (20)$$

If we expand the two terms in Eq. (20) as a pair of Taylor series about $N_r = a$ and then let $a \to \infty$, we get

$$1 - \frac{1}{\log\left(1 - \frac{\mathsf{e}P_{\min}}{N_r}\right)} = \frac{N_r}{\mathsf{e}P_{\min}} + \frac{1}{2} - \frac{\mathsf{e}P_{\min}}{12N_r} - \frac{\mathsf{e}^2P_{\min}^2}{24N_r^2} + O\left(\left(\frac{1}{N_r}\right)^3\right) \quad (21a)$$

$$\begin{aligned} 1 + \frac{\log(P_{\min})}{\log\left(1 - \frac{1}{N_r}\right)} &= -N_r\log(P_{\min}) + \left(1 + \frac{\log(P_{\min})}{2}\right) \\ &\quad + \frac{\log(P_{\min})}{12N_r} + \frac{\log(P_{\min})}{24N_r^2} + O\left(\left(\frac{1}{N_r}\right)^3\right). \end{aligned} \quad (21b)$$

If N_r is large, we can use Eq. (21) to get the following approximate expression for $N_u^{\max}(\theta; N_r)$:

$$N_u^{\max}(\theta; N_r) \approx \begin{cases} \left\lfloor \frac{1}{2} + N_r/(\mathrm{e}P_{\min}) \right\rfloor, & 0 < P_{\min} \leq 1/\mathrm{e} \\ \left\lfloor 1 - \left(N_r - \frac{1}{2}\right) \log(P_{\min}) \right\rfloor, & 1/\mathrm{e} < P_{\min} \leq 1 \end{cases} \tag{22}$$

The approximations are very close to the actual values. In Fig. 3 we plot the approximation error for $\theta^{\max}$ and $N_u^{\max}(\theta; N_r)$ versus N_r. First, in Fig. 3a, we plot the absolute error between the exact value of $\theta^{\max}$ and its approximate value from Eq. (19). The error decreases as N_r increases and $P_{\min}$ decreases. By increasing $P_{\min}$, the error increases to an asymptotic value associated with $P_{\min} = 1/\mathrm{e}$, for any N_r. Beyond $P_{\min} = 1/\mathrm{e}$, $\theta^{\max} = 1$, and the error remains at the level associated with $P_{\min} = 1/\mathrm{e}$. The error is significant for high values of $P_{\min}$ and small pool sizes; for example, it is greater than 0.1 for $N_r < 5$ resources when $P_{\min} \geq 1/\mathrm{e}$. However, discovery resource pools that are so small do not seem likely to be implemented in practical systems. For $N_r > 50$ resources, the greatest error in the value of $\theta^{\max}$ is less than 0.01, and the error is less than 0.001 for $N_r > 500$ resources.

We plot the absolute error in $N_u^{\max}(\theta; N_r)$ versus N_r in Fig. 3b, comparing the exact value from Eq. (10) with the approximation from Eq. (22). We use a different set of values for $P_{\min}$ to show the effect when $P_{\min} > 1/\mathrm{e}$. In this case, the error is less than $1/2$ over the full range of values of N_r. Thus, the approximation in Eq. (22) is accurate enough that it can be used in all cases.

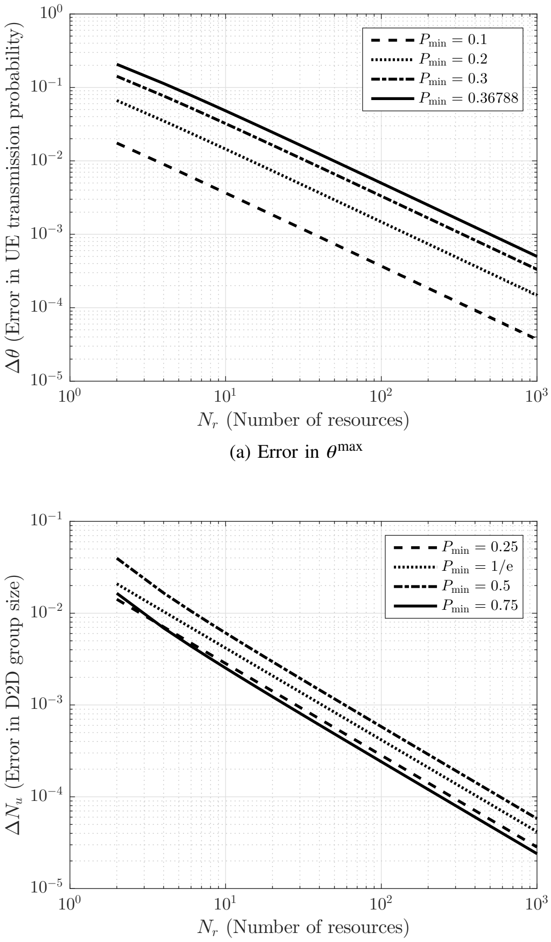
IV. Validation and Numerical Results

We validated our results from Section II and Section III by performing Monte Carlo simulations of discovery pool resource selection in MATLAB, and we compared the resulting maximum group size with what was predicted by Eqs. (3) and (10). We also examined the impact of the discovery pool dimensions on the group size upper bound by performing simulations in ns-3 of a group of OOC UEs exchanging discovery messages over the PSDCH, in which we obtained the maximum group size $N_u^*(\theta; N_r, N_t)$ versus $P_{\min}$, using the following four values for θ: 1/4, 1/2, 3/4, and 1.

A. Monte Carlo Simulations

We simulated the discovery resource selection process for a discovery pool size of $N_r = 40$ resources, using the parameters shown in Table I. As the extent of the resource pool in the time domain decreases, the half-duplex effect has a greater impact on the ability of UEs to decode discovery messages. This will reduce the maximum OOC group size that can achieve $P_{Y \to X} \geq P_{\min}$.

For each set of pool parameters, we used the following three values for $P_{\min}$: 0.2, 0.3, and 0.4. We used Eq. (19) to generate the following respective values of θ: 0.5437, 0.8155, and 1. We assigned the generated value of θ to each UE in the OOC group, and we computed $N_u^*(\theta; N_r, N_t)$ and $N_u^{\max}(\theta; N_r)$, which are listed in Table I.



(a) Error in $\theta^{\max}$

(b) Error in $N_u^{\max}(\theta; N_r)$

Fig. 3. Errors between approximate and exact values of $\theta^{\max}$ and $N_u^{\max}(\theta; N_r)$

TABLE I
Monte Carlo Simulation Parameters

N_f	N_t	$P_{\min}$	θ	N_u^*	$N_u^{\max}$
1	40	0.2	0.5437	74	74
		0.3	0.8155	49	49
		0.4	1.0000	37	37
4	10	0.2	0.5437	70	74
		0.3	0.8155	46	49
		0.4	1.0000	34	37
10	4	0.2	0.5437	64	74
		0.3	0.8155	39	49
		0.4	1.0000	26	37

For each set of pool parameters and value of $P_{\min}$, we simulated N_u UEs, where N_u varied from $N_u^* - 2$ to $N_u^{\max} + 2$, because we are interested in examining the behavior of $P_{Y \to X}$ in the vicinity of the theoretical upper bounds on N_u. For each value of N_u we performed $N_{\text{runs}} = 100$ runs, with $N_{\text{trials}} = 1000$ trials per run. For each trial, we designated a random UE as UE_0, the transmitter of interest, and we chose a second random UE to be the receiver. Each UE independently generated a random transmission decision variable, $p1$. If $p1 \leq \theta$, the UE transmitted; otherwise, it did not. UEs that decided to transmit discovery messages chose one of the N_r resources in the pool with uniform probability. Each UE's resource choice was independent of the choices of all other UEs. UEs kept track of which subframe set they were using to send their discovery messages. The trial was a success if UE_0 decided to transmit, no other UE chose UE_0's resource, and the receiver did not transmit in the same subframe set as UE_0. Any other outcome was a failure.

The estimated value of $P_{Y\to X}$ for each run is the ratio of the number of successful trials to the number of trials in the run. The output of each simulation was a set of N_{runs} estimates for $P_{Y\to X}$: $\{\widehat{P}_{Y\to X}(n)\}_{n=1}^{N_{\text{runs}}}$. The estimate of $P_{Y\to X}$ is

$$\widehat{P}_{Y\to X} = \frac{1}{N_{\text{runs}}} \sum_{n=1}^{N_{\text{runs}}} \widehat{P}_{Y\to X}(n) \qquad (23)$$

The standard deviation of the set of estimates of $P_{Y\to X}$ is:

$$\widehat{\sigma}_{\widehat{P}_{Y\to X}} = \sqrt{\frac{1}{N_{\text{runs}}-1} \sum_{n=1}^{N_{\text{runs}}} \left(\widehat{P}_{Y\to X}(n) - \widehat{P}_{Y\to X}\right)^2} \qquad (24)$$

Using Eqs. (23) and (24), we plot the 95 % confidence interval associated with the simulation result, whose limits are

$$\widehat{P}_{Y\to X} \pm 1.96\,\widehat{\sigma}_{\widehat{P}_{Y\to X}}/\sqrt{N_{\text{runs}}}. \qquad (25)$$

Fig. 4 shows the simulation results that we obtained using the parameters listed in Table I. In each subfigure, we plot $P_{Y\to X}$ versus N_u for a given set of pool parameters, with different markers for each of the three values of $P_{\min}$ that we used. We show the theoretical values of N_u^*, listed in Table I, in each subfigure with additional ticks on the N_u-axis. In each subfigure, $P_{Y\to X}$ decreases with respect to N_u, so that N_u^* is the largest value of N_u such that $P_{Y\to X} \geq P_{\min}$, which validates the model. The slope of $P_{Y\to X}$ decreases as $P_{\min}$ decreases, so that $P_{Y\to X}$ becomes less sensitive to N_u. Note that in Fig. 4a, $N_u^* = N_u^{\max}$. Also, the sensitivity of N_u^* to $P_{\min}$ increases as $P_{\min}$ becomes small.

Figs. 4b and 4c show the effect of modifying the discovery resource pool parameters while keeping N_r constant. For a pool with $N_f = 4$ PRB pairs and $N_t = 10$ subframe sets, N_u^* is below $N_u^{\max}$ by 5.71 %, 6.52 %, and 8.82 % for $P_{\min}$ values of 0.2, 0.3, and 0.4, respectively. The effect is greater when $N_f = 10$ PRBs and $N_t = 4$ subframe sets, where N_u^* is below $N_u^{\max}$ by 15.63 %, 25.64 %, and 42.31 % for $P_{\min}$ values of 0.2, 0.3, and 0.4, respectively.

B. Simulations using ns-3

Our ns-3 simulations examine the impact of the pool parameters on the group size upper bound. We examined four values of θ: 0.25, 0.50, 0.75, and 1.00. For each value of θ, we created groups of UEs whose sizes were multiples of 15, up to a maximum of 150 UEs, and we examined groups consisting of 2 UEs. All UEs used the same value of θ. For each combination of group size and θ, we considered the same three discovery resource pool configurations that we examined in Section IV-A, each of which has $N_r = 40$ resources.

For each set of parameters, we conducted 10 runs, with 10 trials per run. We obtained estimates of $P_{Y\to X}$ for each trial, and obtained estimates of the mean and standard deviation using Eqs. (23) and (24), and used Eq. (25) to obtain 95 % confidence intervals. Fig. 5 shows the results; we use connecting lines to make the graphs legible and to illustrate trends in the data.

We compared our results to the theoretical model and found excellent agreement; we do not show the theoretical results in Fig. 5 for the sake of clarity in the subfigures. The figures collectively show that the group size upper bounds are most sensitive to $P_{\min}$ when θ is large, since increasing each UE's transmission rate increases the offered load, which increases the collision rate. Also, the effect of the pool parameters on the upper bound is more pronounced when $P_{\min}$ is small and when θ is large. The plots show that if we want $P_{\min}$ to be close to θ (e.g., within 90 %), N_u must be small. For example, when $\theta = 0.25$, we can satisfy $P_{\min} \approx 0.9\,\theta$ with a group that comprises at most about 15 UEs. Larger upper bounds for the group size are achievable only with a larger resource pool, preferably one where $N_t > N_f$.

V. Summary

In this paper, we derived an upper bound for the size of a group of OOC UEs whose members' probability of decoding discovery messages is at least $P_{\min}$. We showed that the value of θ that maximizes the group size upper bound is insensitive to the discovery pool dimensions. By using a linear approximation for the upper bound, we developed a closed-form expression for the optimal value of θ. We then showed that the optimal value of θ is proportional to $P_{\min}$, which yields a simple expression for the maximum group size. We validated our results using both Monte Carlo simulations in MATLAB and simulations of the PSDCH used by OOC UEs in the ns-3 network simulation tool.

References

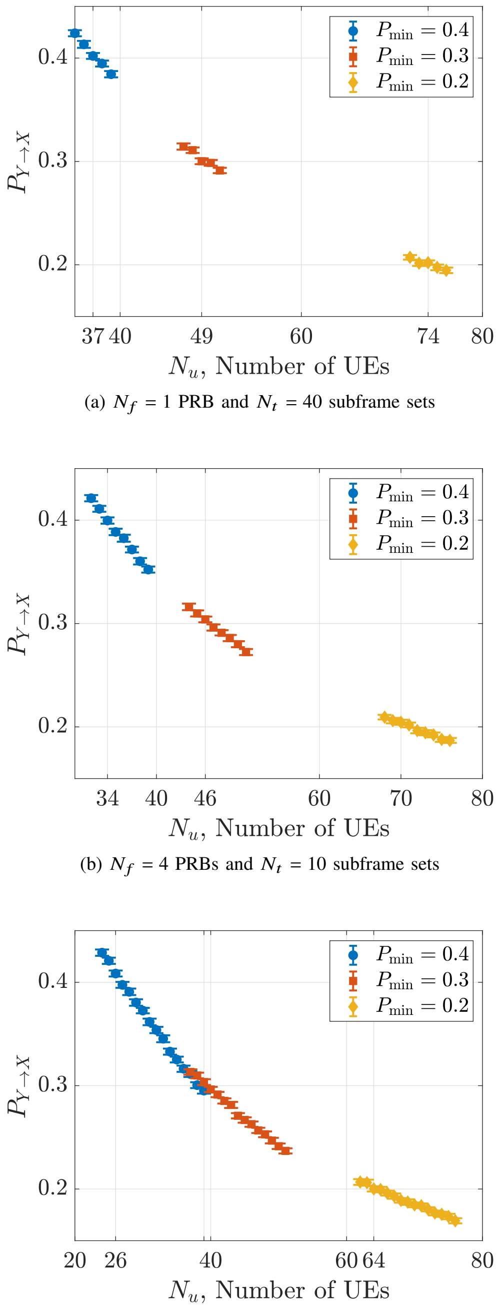
[1] 3GPP, "Feasibility study for Proximity Services (ProSe)," 3rd Generation Partnership Project (3GPP), TR 22.803 V12.2.0, June 2013. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/22_series/22.803/22803-c20.zip

[2] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), TS 36.321 V12.7.0, September 2015. [Online]. Available: http://www.3gpp.org/ftp/Specs/archive/36_series/36.321/36321-c70.zip

[3] D. Griffith and F. Lyons, "Optimizing the UE transmission probability for D2D direct discovery," in *2016 IEEE Global Telecommunications Conference (GLOBECOM 2016)*, December 2016.

[4] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert *W* function," *Advances in Computational Mathematics*, vol. 5, no. 1, pp. 329–359, 1996.

Griffith, David W.; Ben Mosbah, Aziza. "On Upper Bounds for D2D Group Size." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 13, 2019.

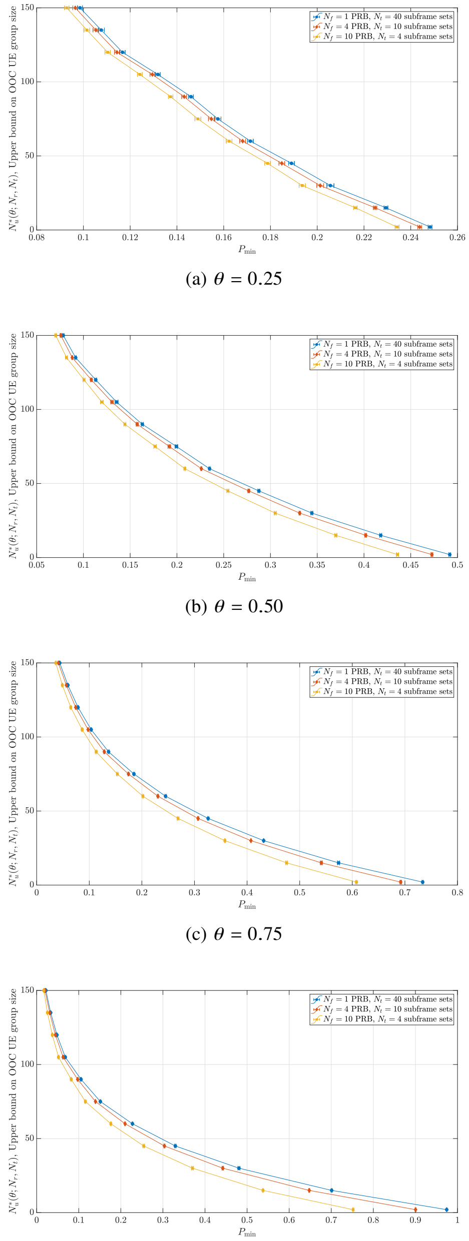


(a) N_f = 1 PRB and N_t = 40 subframe sets

(b) N_f = 4 PRBs and N_t = 10 subframe sets

(c) N_f = 10 PRBs and N_t = 4 subframe sets

Fig. 4. Monte Carlo simulation results showing estimates of $P_{Y \to X}$ versus OOC UE group size N_u, for various pool size parameters and values of $P_{\min}$, with 95 % confidence intervals shown.



(a) θ = 0.25

(b) θ = 0.50

(c) θ = 0.75

(d) θ = 1.00

Fig. 5. Simulation results from ns-3 showing estimates of $N_u * (\theta; N_r, N_t)$ versus $P_{\min}$, for various values of θ, with 95 % confidence intervals shown.

AlGaN/GaN core-shell heterostructures for nanowire UV LEDs

Matt Brubaker[1], Bryan Spann[1], Kristen Genter[1,2], Alexana Roshko[1], Paul Blanchard[1], Todd Harvey[1], Kris Bertness[1]

[1] Physical Measurement Laboratory, National Institute of Standards and Technology, Boulder, CO
[2] Department of Mechanical Engineering, University of Colorado, Boulder, CO

Email: matthew.brubaker@nist.gov

Keywords: GaN nanowires, UV LEDs, core-shell

Nanowire-based ultraviolet (UV) LEDs hold great promise as nanoscale light sources, potentially enabling advanced scanning microscopy probes capable of optoelectronic sensing and near-field scanning photolithography. In this work, we report on the synthesis of core-shell nanowire LEDs, including characterization of *p-i-n* junctions and AlGaN/GaN heterostructures for enhanced injection and confinement of minority carriers.

Ordered nanowire arrays were grown on silicon substrates via an N-polar selective area nanowire growth process.[1] These silicon-doped nanowire cores were then overcoated with a conformal Mg:GaN shell layer to produce nanowires arrays with *p-i-n* structure. As shown in Figure 1a, the nanowire LEDs produce electroluminescence under forward bias with emission at 380 nm, corresponding to donor-acceptor-pair (DAP) recombination in the *p*-type shell layer. For extending the LED emission towards shorter wavelengths, nanowire arrays with AlGaN shell layers were also grown using conditions identical to those used for the Mg:GaN shells. The AlGaN overcoats were found to exhibit thickness and compositional gradients that depended on the nanowire diameter. As shown in Figure 1b, the thickness of the axial segment of the AlGaN shell was larger in small-diameter nanowires. The Al mole fraction was also observed to decrease for small-diameter nanowires, as determined by photoluminescence measurements (Figure 1c). These observations are consistent with diffusion-induced transport of Ga (but not Al) to the nanowire tip during growth, which effectively increased the axial growth rate and diluted the alloy concentration at the tip.

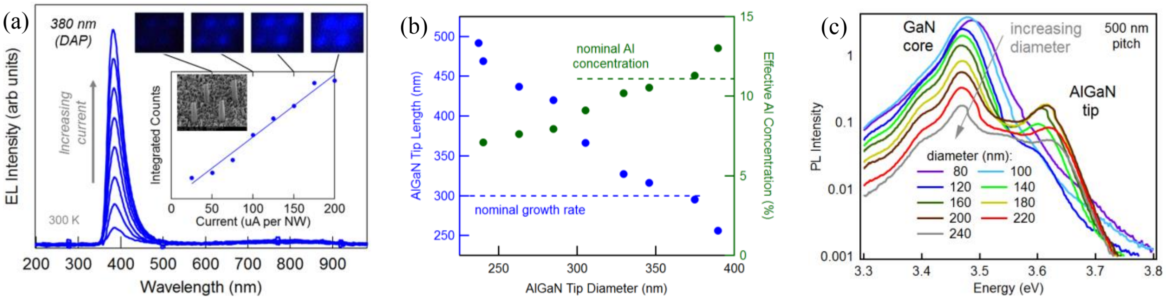


Figure 1. *(a) Core-shell p-i-n GaN nanowire LEDs emitting at 380 nm and (b) dimensional and (c) photoluminescence measurements of AlGaN/GaN core-shell heterostructures. The dotted lines in (b) indicate equivalent planar values for the nominal growth rate and Al mole fraction.*

References

[1] M. Brubaker, *et al*, *Cryst. Growth & Des.* **16** (2015), 596.
[2] Contribution of an agency of the U.S. government; not subject to copyright.

Estimating Interzonal Leakage in a Net-Zero Energy House

Lisa Ng[1], Lindsey Kinser[2], Steven Emmerich[1], Andrew Persily[1]

[1] Indoor Air Quality and Ventilation Group, Engineering Laboratory, National Institute of Standards and Technology, Gaithersburg MD
[2] College of Engineering, Mechanical Engineering and Mechanics, Drexel University

ABSTRACT

The Net-Zero Energy Residential Test Facility (NZERTF) was constructed at the National Institute of Standards and Technology (NIST) to support the development and adoption of cost-effective net-zero energy designs and technologies. The 250 m^2 two-story, unoccupied NZERTF, built in 2012, had among its design goals an airtight and highly insulated building enclosure designed for heat, air and moisture control. The airtightness goal was achieved through detailed envelope design, and careful construction, as well as during and after construction commissioning. When it was built, the NZERTF was one of the tightest residential buildings in North America with a whole building pressurization test result of roughly 0.6 h^{-1} at 50 Pa measured per ASTM E779-19, *Standard Test Method for Determining Air Leakage Rate by Fan Pressurization*. No special attention was given to the airtightness of the interior floors and other interior partitions. To support airflow modeling efforts, this interior leakage was quantified through a series of interzonal pressurization tests. Both the basement and attic were considered to be conditioned spaces because the thermal and air-moisture barriers encompass the basement walls and attic roof. Transfer grilles and other openings linked the living space to these two zones. A series of fan and partition configurations were used to quantify the leakage values of the various interzone airflow paths. Test results showed that the interior floors were more than 10 times leakier than the exterior building envelope and that the leakage associated with the transfer grilles between levels was less than the floor leakage. This paper describes the design of the interzonal tests and the challenges in

Ng, Lisa; Kinser, Lindsey; Emmerich, Steven; Persily, Andrew K. "Estimating Interzonal Leakage in a Net-Zero Energy House." Paper presented at ASTM Symposium on Whole Building Air Leakage: Testing and Building Performance Impacts, San Diego, CA, US. April 08, 2018 - April 09, 2018.

performing them, which included isolating zones, controlling multiple blower doors, and access for installing pressurization fans. The results of these tests were inputs to a multizone airflow (CONTAM) model of the building for use in evaluating the effects of different ventilation strategies and other airflow-related technologies on energy consumption and indoor air quality.

Keywords

airtightness, interzonal leakage, net-zero house, ASTM E779-10, pressurization tests

INTRODUCTION

In 2017, buildings were associated with 39 % of all energy used in the United States, with residential buildings and commercial buildings accounting for 20 % and 19 %, respectively [1]. Based on estimates by the U. S. Department of Energy (DOE), infiltration alone accounts for 14 % and 6 % of the energy used by residential and commercial buildings, respectively [2]. To reduce these energy impacts, tighter building envelopes are being required by codes and standards [2-5]. There are no standards related to the leakiness of interior floors and walls based on energy and indoor air quality considerations (though some fire codes address the issue), even though interzonal airflow through these surfaces can be important for contaminant transport and thus occupant exposure [6].

The literature on interzonal airflow experiments includes analyses of two-zone test cases [7-9]. Determining interzonal airflow is more difficult than whole building testing, either requiring a series of pressurization tests or multiple tracer gas tests. Emmerich et al. [7] conducted interzonal pressurization tests in five homes with attached garages to determine the leakiness of the house-garage interface. The tests involved placing one or two fans in different exterior doorways and altering the positions of the door connecting the two zones, as well as of the garage door to the outside. They found that the house-garage interface (normalized by the house-garage surface area) was two and half times to nearly eleven times leakier than the house exterior envelope (normalized by the house exterior surface area), which has implications for the transfer of contaminants from garages into houses. Hult et al. [9] compared the results of various single-fan and two-fan tests of the leakiness of house-garage interfaces. They also compared single-pressure difference to multiple-pressure difference tests. They found that a method using one fan in two configurations provided results with the smallest uncertainty among the single-fan tests. In general, the single-pressure tests led to less reliable results than the multiple-pressure tests. Though requiring an extra configuration, the "one-fan, three configurations" test performed by Emmerich et al. [7] also resulted in low uncertainty.

Tracer gas tests have also been used to determine interzonal airflow rates, as opposed to the interzone partition leakiness determined by the previously-described fan pressurization tests [10-15]. Du et al. [16] conducted constant-concentration tracer tests in several homes using two tracer gases. The steady-state concentrations of the tracer gases were used to estimate the airflow rates between a bedroom and the rest of the home. They found that most of the air entering the bedrooms came from somewhere else in the house and not from outside. Conversely, most of the air entering the rest of the house came from outside and not from the bedroom. They also found that homes that relied more on central heating and cooling systems had relatively higher interzonal airflows than homes in which occupants opened windows for ventilation.

This paper examines the use of blower door tests to determine the effective leakage area (ELA) of various house components (exterior envelope, interior floor leakage, leakage of transfer grilles) for input into a multizone airflow model. The results of such simulations can be used to evaluate different ventilation strategies and other airflow-related technologies to study the effects of weather, indoor conditions, and system operation on interzonal airflow and contaminant transport. This paper describes the test house, the design of the interzonal tests and the challenges in performing these tests, as well as the performance of a multizone airflow model (i.e., CONTAM) using measured values of interior leakages to predict contaminant concentrations. Interzonal tracer decay tests also were performed, and their analyses are saved for future work.

TEST HOUSE

The Net-Zero Energy Residential Test Facility (NZERTF) was built on the campus of the National Institute of Standards and Technology (NIST) in 2012 to demonstrate low-energy residential technologies with the goal of net-zero energy use on an annual basis (FIG. 1). The NZERTF is a 250 m^2 two-story, unoccupied house located in Gaithersburg, MD with an unfinished basement and an attic, both within the conditioned space. As reflected in TABLE 1, which summarizes the physical characteristics of the house, the basement is mostly below-grade, with a window well providing egress. A two-story foyer with a staircase connects the first and second floor, which has a horizontal area of approximately 17 m^2 (12 % of the first floor area). Given the open connection between these two floors, they were considered a single zone in these tests.

The detached garage contains the controls and data acquisition systems of the instruments and sensors in the NZERTF, so that their heat load is not introduced into the home. Lighting, appliances, plug loads, and sensible and latent heat loads of the simulated occupants are controlled by the data acquisition system [17]. A virtual family consisting of two adults and two children are simulated in the house, with their electrical and water usage varying over a seven-day schedule [18].

TABLE 1. Physical characteristics of NZERTF

Building characteristic	**Value**
Roof area	184 m^2
Basement wall area (above-grade)	2 m^2
First floor/second floor exterior wall area	314 m^2
Attic floor	130 m^2
Basement ceiling	151 m^2
Total exterior surface area	500 m^2
Total volume (basement, first floor, second floor, attic)	1300 m^3

The main design goal of the NZERTF was to achieve net-zero energy use over the course of a year, which was achieved from July 2013 to June 2014. One of the ways to reduce energy use in homes is to reduce heating and cooling loads because Since infiltration can account for 14 % of the total energy use of a home, the NZERTF was designed and constructed to be airtight. The building envelope airtightness of the NZERTF was tested to be 0.63 h^{-1} at 50 Pa [19], which is tighter than the requirements in LEED v4 [20] and ENERGY STAR v3.1 [21], and only slightly leakier than the Passive House U. S. requirement [22]. The normalized leakage (NL) value for the house equals 0.06, which is tighter than 99 % of U.S. homes based on statistical analysis of the Lawrence Berkeley National Laboratory Residential Diagnostics Database [23]. The NL value is defined in the ASHRAE Fundamentals Handbook [24] as follows:

$$NL = 1000 \left(\frac{ELA_4}{Area}\right)\left(\frac{H}{2.5\ m}\right)^{0.3} \tag{1}$$

where ELA_4 (m^2) is the ELA at 4 Pa, Area (m^2) is the floor area, and H (m) is the house height.

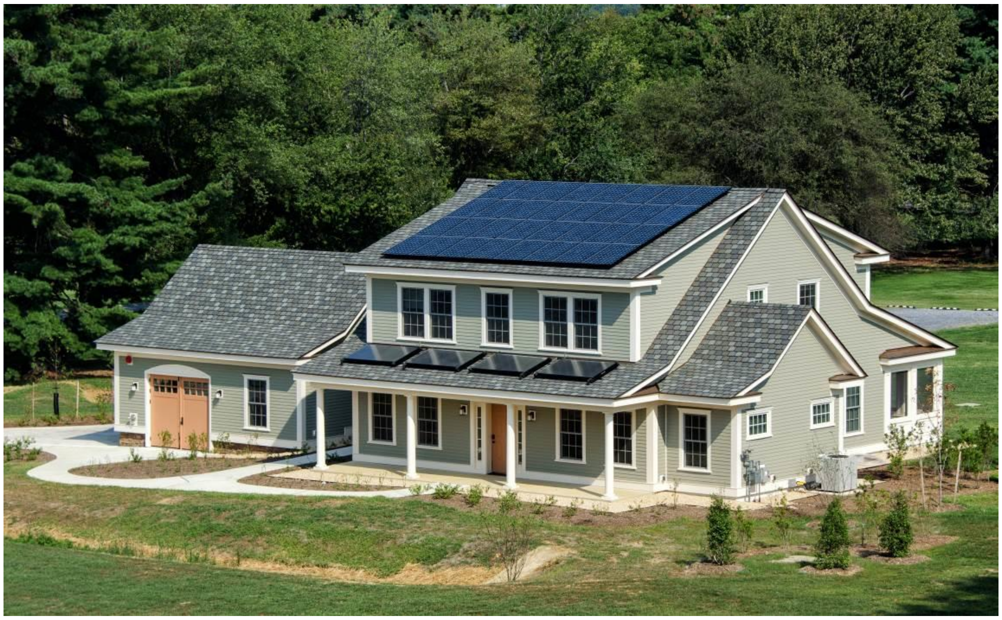


FIG. 1. NZERTF at NIST facing south.

The basement, first floor, and second floor were actively conditioned by a central, air-source heat pump. The central heat pump had supplies in the basement, first floor, and second floor. Two returns were on the first floor and two returns were on the second floor. The heat pump has no outdoor air intake. Three transfer grilles were located on the floor of the first level to allow airflow between the basement and the house (referred to as the first floor transfer grilles). Two transfer grilles were located on the ceiling of the second floor to allow airflow between the attic and the house (FIG. 2a) referred to as the attic transfer grilles. Because the attic is within the thermal envelope, the attic transfer grilles were installed to provide the attic with conditioned air without requiring air distribution ductwork in the attic [25]. All the transfer grilles contain a damper that would close in case of a fire. The basement door and the attic hatch were closed during normal operation of the house. The basement door had an undercut that was approximately 0.9 m wide by

Ng, Lisa; Kinser, Lindsey; Emmerich, Steven; Persily, Andrew K. "Estimating Interzonal Leakage in a Net-Zero Energy House." Paper presented at ASTM Symposium on Whole Building Air Leakage: Testing and Building Performance Impacts, San Diego, CA, US. April 08, 2018 - April 09, 2018.

2.5 cm high, and the attic hatch door had a gap around it that was approximately 0.3 cm wide and 5.5 m long around its perimeter. Under depressurization, smoke tests were performed around interior airflow paths, such as the transfer grilles and around the basement door (FIG. 2) to better understand the airflow through these building elements. Smoke tests were also used to identify airflow paths that should be sealed during pressurization testing, such as the access panels in a bathroom used for signal and control wire (FIG. 2c) to isolate and determine leakage associated with the construction of the floors, basement door, and attic door. Other interior airflow paths that were sealed during testing included the supplies and returns of the central heat pump and the independently ducted mechanical ventilation system, the heat recovery ventilator (HRV). The HRV was balanced, with supplies on the first and second floors and exhausts in the bathrooms on the first and second floors. Both systems were turned off during all of the pressurization tests. The kitchen exhaust fan and dryer also were turned off, and their exterior vents were sealed.

(a) Floor transfer grille (b) Basement door (c) Access panel

FIG. 2 Photographs of smoke tests performed at NZERTF at the (a) floor transfer grilles, (b) basement door, and (c) access panel while depressurizing. Arrows indicate direction of airflow.

Because the NZERTF is airtight, the designers wanted to prevent the house from depressurizing when either the kitchen exhaust fan or dryer were turned on. Thus, a 15-cm round duct was installed in the attic, penetrating the exterior attic wall on the west side, with motorized and barometric dampers installed in the duct. The motorized damper was activated when either the kitchen exhaust fan or dryer was turned on. The barometric damper would open if the motorized damper was open and if the inside pressure was 10 Pa less than the outside pressure.

METHODOLOGY

The testing and data analysis methodologies as follows:

(1) Five house configurations were tested under various blower door arrangements.

(2) Each test configuration was expressed in mathematical form following the analogy of an electrical circuit with pressure corresponding to voltage and airflow corresponding to current. The airflow Q and pressure difference ΔP across a surface were represented by the equation $Q = C\Delta P^{n}$, where C is the flow coefficient ($m^3/s{\bullet}Pa$) and n is the pressure exponent. Both C and n were determined from the test data. These expressions formed the system of the equations needed to solve for the ELAs of the following building surfaces: roof (L_{R}), basement wall (L_{BW}), living area (first floor/second floor) walls (L_{W}), attic floor (L_{AF}), and basement ceiling leakage (L_{BC}) (illustrated in FIG. 3).

(3) The simultaneous solution of this system of equations resulted in values of C for each of the list building surfaces. C was converted to effective leakage, L, using the following equation

[24]:

$$L\ (\text{cm}^2)\ \text{at}\ \Delta P_{\text{ref}} = C \bullet (\rho/2)^{0.5} \bullet (\Delta P_{\text{ref}})^{\text{n-0.5}} \bullet 100^2 \qquad (2)$$

where ΔP_{ref} is the reference differential pressure (Pa) and ρ is the density of air (kg/m^3).

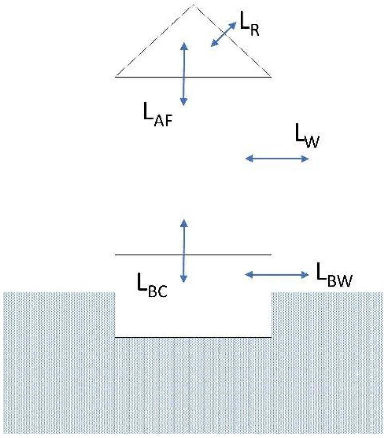


FIG. 3. Building surface leakages in NZERTF.

(4) The effective leakage of the basement door undercut and transfer grilles were determined by subtracting the result of the comparable "sealed" test configuration from the "unsealed" test result.

(1) Test Setup

Two blower doors, which complied with the requirements of ASTM E779-19, *Standard Test Method for Determining Air Leakage Rate by Fan Pressurization,* were used in the tests. Pressures were measured using a multichannel pressure measurement and fan control apparatus from the same manufacturer as the blower doors. Baseline pressures were recorded when the fan was turned off before and after each blower door test. These values were then averaged and subtracted from each measured pressure recorded during the test, as outlined in ASTM E779. The data logging and control software allowed for the simultaneous measurement of differential pressures across the fan and at three locations throughout the house. The software was setup to record differential pressures and fan flow rates between 10 Pa and 60 Pa in increments of 5 Pa, in both pressurization and depressurization modes. The software also provided the ability to take 100 pressure differential readings at each incremental differential pressure value and report the average. All tests were conducted over two weeks in November 2016, during which time the average indoor temperature was 21 °C, outdoor temperature was 7 °C and wind speed was 4 m/s.

(2) Test Configurations

Five house configurations were tested, varying the placement of the fan and open/closed status of exterior and interior doors (FIG. 4). All tests were successfully executed except for Test #3, which required the attic pressure to be at the same as the outside pressure. The existing opening in the attic (makeup-air duct) was not large enough to neutralize the attic-outdoor pressure. Details of each test and the mathematical expressions used to represent each test are described in the following sections.

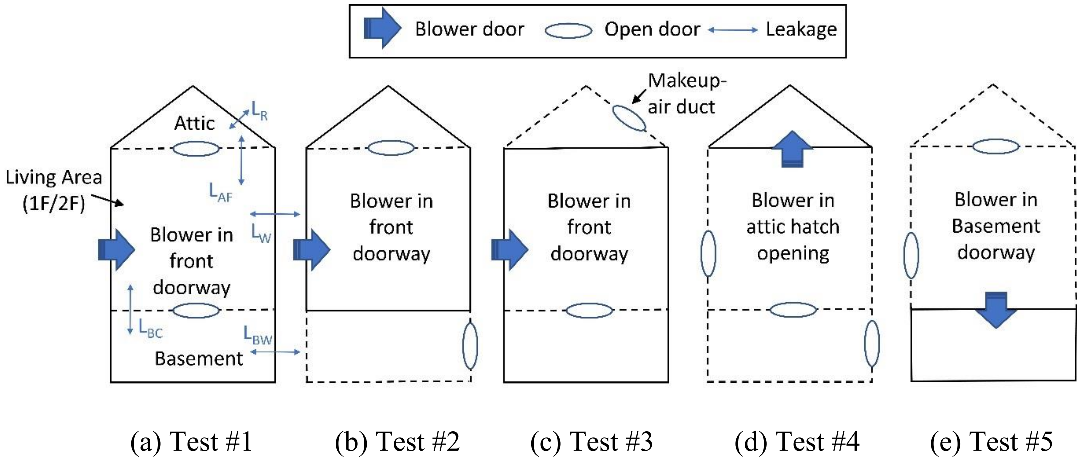


FIG. 4. Five house configurations for determining external and interzonal leakage.

Test #1

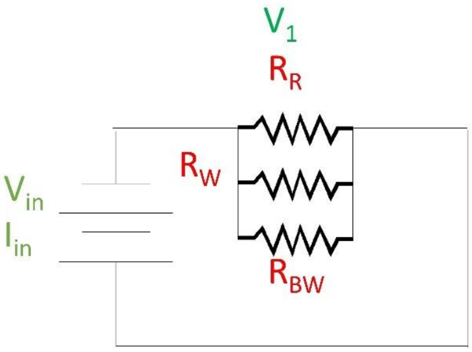
For this test configuration, the blower door was placed in the front doorway of the house, and both the basement door and attic hatch remained open. Pressure differentials with respect to the outdoors were measured in the basement, attic, and living room to ensure that the induced pressure across the building envelope was uniform across the entire house. Two subconfigurations ("A" and "B") were also tested: in Test #1A the exterior dryer and kitchen exhaust vents were unsealed and in Test #1B, these vents were sealed.

Test #1 was represented as an electrical circuit with three "resistances" (analogous to the flow coefficient C) in parallel (FIG. 5). The sum of the flow through each branch of the "circuit" is equal to Q_1, which was the measured airflow delivered by the blower door fan test to yield values of C_1 and n_1:

$$Q_1 = C_1 \Delta P_1^{n_1} = C_R \Delta P_{R_1}^{n_1} + C_W \Delta P_{W_1}^{n_1} + C_{BW} \Delta P_{BW_1}^{n_1} \tag{3}$$

Note that the value of n_1 is assumed to be the same for all surfaces in Eq. (3). No tests were performed to determine these values of n individually as part of Test #1. The subscript 1 denotes Test #1, and subscripts R, BW, and W denote the building surfaces (roof, basement wall and walls, respectively). On the basis of consideration of the similarity in airtightness of the exterior surfaces throughout the house, and the inability to measure them separately, we assumed that the following relationship:

$$C_{\mathrm{R}}/A_{\mathrm{roof}} = C_{\mathrm{BW}}/A_{\mathrm{basement\ wall}} = C_{\mathrm{W}}/A_{\mathrm{walls}} \tag{4}$$



I_{in}: current in (analogous to ΔP of test)
V_{in}: voltage in (analogous to Q of test)
V_1: voltage through parallel resistances R_{R}, R_{BW} and R_{W}
R_{R}: resistance through roof (analogous to C_{R})
R_{BW}: resistance through basement wall (analogous to C_{BW})
R_{W}: resistance through walls (analogous to C_{W})

FIG. 5. Electrical circuit equivalent of Test #1

Test #2

For this test configuration, the blower door was placed in the front doorway. The basement door was closed and the attic hatch was opened. Four subconfigurations (A, B, C, and D) were tested, varying the sealing and unsealing of the first floor transfer grilles (between the first floor and the basement) and the sealing and unsealing of the basement door undercut as summarized in Table 2.

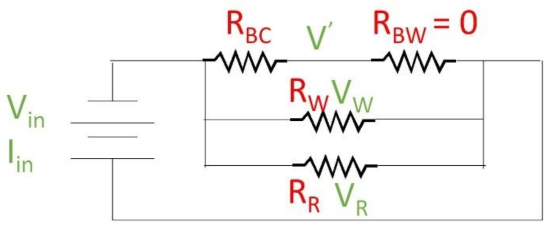
TABLE 2. Summary of subconfigurations for Test #2

Test #2 subconfigurations	First floor transfer grilles	Basement door undercut
A	Unsealed	Unsealed
B	Sealed	Unsealed
C	Unsealed	Sealed
D	Sealed	Sealed

Test #2 also has three resistances in parallel (C_R, C_W, C_2), with one of the resistances ($C^{'}$) composed of two resistances in series (C_{BW}, C_{BC}) (FIG. 6). Because the basement window is open for Test #2, the resistance R_B is essentially zero. The sum of the airflow through each branch of the "circuit" is equal to Q_2, which was measured during the blower door tests and yielded values of C_2 and n_2. The airflow-pressure expressions are as follows, where the subscript 2 refers to Test #2. Note that the value of n_2 is assumed to be the same for all surfaces in Eqs. (5) and (6).

$$Q_2 = C_2 \Delta P_2^{n_2} = C^{'} (\Delta P^{'})^{n^{'}} + C_W \Delta P_{W_2}^{n_2} + C_R \Delta P_{R_2}^{n_2}, \text{ and} \tag{5}$$

$$C'(\Delta P')^{n'} = C_{BC} \Delta P_{BC_2}^{n_2} \tag{6}$$



I_{in}: current in (analogous to ΔP of test)
V_{in}: voltage in (analogous to Q of test)
V_{R}: voltage through R_{R}
V_{W}: voltage through R_{W}
V': voltage through combined R_{BW} and R_{BC}

R_{R}: resistance through roof (analogous to C_{R})
R_{BW}: resistance through basement wall (analogous to C_{BW})
R_{W}: resistance through walls (analogous to C_{W})
R_{BC}: resistance through basement ceiling (analogous to C_{BC})

FIG. 6. Electrical circuit equivalent of Test #2.

Test #3

In Test #3, the blower door was placed in the front doorway. The basement door was open and the attic hatch was closed. Two subconfigurations (A and B) were tested: in Test #3A the attic transfer grilles (between the second floor and attic) were unsealed and in Test #3B, these transfer grilles were sealed. The airflow-pressure expressions are as follows:

$$Q_3 = C_3 \Delta P_3^{n_3} = C^{'}\left(\Delta P^{'}\right)^{n^{'}} + C_W \Delta P_{W_3}^{n_3} + C_{BW} \Delta P_{BW_3}^{n_3} \tag{7}$$

$$C'(\Delta P')^{n'} = C_R \Delta P_{R_3}^{n_3} = C_{AF} \Delta P_{AF_3}^{n_3} \tag{8}$$

Note that the value of n_3 is assumed to be the same for all surfaces in Eqs. (7) and (8).

Test #4

In Test #4, a smaller fan designed for duct leakage tests was connected to the attic hatch opening using a plywood mount and flexible duct (FIG. 7), and the front door of the house was open. Four subconfigurations (A, B, C, and D) were tested, alternating sealing and unsealing of the attic transfer grilles (between the second floor and the attic) and makeup-air duct in the attic (Table 3).

The airflow-pressure expression is as follows:

$$Q_4 = C_4 \Delta P_4^{n_4} = C_R \Delta P_{R_4}^{n_4} + C_{AF} \Delta P_{AF_4}^{n_4} \tag{9}$$

Note that the value of n_4 is assumed to be the same for all surfaces in Eq. (9).

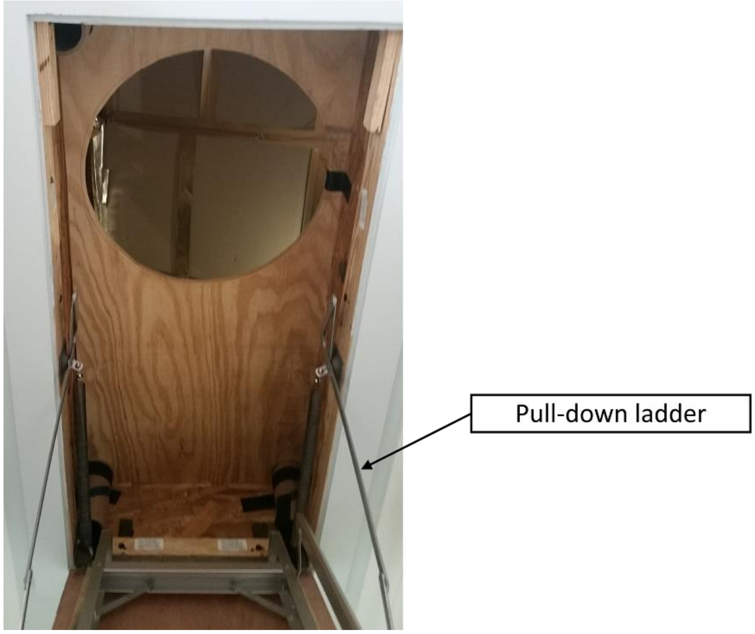


FIG. 7. Plywood mount for smaller fan in attic hatch opening.

TABLE 3. Summary of subconfigurations for Test #4.

Test #4 subconfigurations	Attic transfer grilles	Makeup-air duct
A	Sealed	Unsealed
B	Unsealed	Unsealed
C	Sealed	Sealed
D	Unsealed	Sealed

Test #5

For Test #5, the blower door was placed in the basement doorway and the front door of the house was open. Two subconfigurations (A and B) were tested: in Test #5A the first floor transfer grilles were sealed and in Test #5B, these transfer grilles were unsealed. The airflow-pressure expression is as follows:

$$Q_5 = C_5 \Delta P_5^{n_5} = C_{BW} \Delta P_{BW_5}^{n_5} + C_{BC} \Delta P_{BC_5}^{n_5} \tag{10}$$

Note that the value of n_5 is assumed to be the same for all surfaces in Eq. (10).

(3) Determining Leakage Values of Building Surfaces

This section describes the determination of the individual leakage values from the results of the tests that were just described, solving for C_{R}, C_{BW}, C_{W}, C_{AF}, and C_{BC}. Using the assumption expressed in Eq. (4), the only unknowns were C_{AF} and C_{BC}. They were solved, respectively, using Eq. (9) for Test #4C and Eq. (10) for Test #5A. These test numbers were the subconfigurations in which the interior leakage paths (basement door undercut and transfer grilles) were sealed. The values of C_{R}, C_{BW}, C_{W}, C_{AF} and C_{BW} were then converted to effective leakages L using Eq. (2).

The values of L_{AF} and L_{BC} determined using C_{AF} and C_{BC}, respectively, were compared with values calculated by subtracting the result of the comparable "sealed" test configuration from the "unsealed" test result. The ELA of Test #5A (first floor transfer grilles sealed) minus the leakage area of the basement wall (L_{BW}) equals the leakage area of the basement ceiling (L_{BC}). The ELA of Test #4C (attic transfer grilles sealed) minus the leakage area of the roof (L_{R}) equals the leakage area of the attic floor (L_{AF}).

(4) Determining Leakage Values of Basement Door Undercut and Transfer Grilles

The leakages of the basement door undercut and transfer grilles were determined by subtracting the result of the comparable "sealed" test configuration from the "unsealed" test result. The solution process for each building component is described below:

The effective leakage of the basement door undercut was calculated two ways:

(1) ELA Test #2 with first floor transfer grilles unsealed: $\mathrm{ELA}_{2\mathrm{A}}$ (door undercut unsealed) minus $\mathrm{ELA}_{2\mathrm{C}}$ (door undercut sealed)

(2) ELA Test #2 with first floor transfer grilles sealed: $\mathrm{ELA}_{2\mathrm{B}}$ (door undercut unsealed) minus $\mathrm{ELA}_{2\mathrm{D}}$ (door undercut sealed).

The effective leakage of the first floor transfer grilles was calculated three ways:

(1) ELA Test #2 basement door undercut unsealed: $\mathrm{ELA}_{2\mathrm{A}}$ (transfer grilles unsealed) minus $\mathrm{ELA}_{2\mathrm{B}}$ (transfer grilled sealed)

(2) ELA Test #2 basement door undercut sealed: $\mathrm{ELA}_{2\mathrm{C}}$ (transfer grilles unsealed) minus $\mathrm{ELA}_{2\mathrm{D}}$ (transfer grilled sealed)

(3) $\mathrm{ELA}_{5\mathrm{A}}$ (transfer grilles unsealed) minus $\mathrm{ELA}_{5\mathrm{B}}$ (transfer grilles sealed)

The effective leakage of the attic transfer grilles was calculated two ways:

(1) ELA Test #4 makeup-air duct sealed: $\mathrm{ELA}_{4\mathrm{A}}$ (transfer grilles unsealed) minus $\mathrm{ELA}_{4\mathrm{B}}$ (transfer grilled sealed)

(2) ELA Test #4 makeup-air duct unsealed: $\mathrm{ELA}_{4\mathrm{C}}$ (transfer grilles unsealed) minus $\mathrm{ELA}_{4\mathrm{D}}$

(transfer grilled sealed).

RESULTS

This section summarizes the ELAs obtained from the 14 blower door tests. These ELAs are used to calculate the leakiness of the roof, first and second floor walls, basement walls, basement ceiling, and attic floor using derived flow coefficients. The test ELAs also are used to calculate the effective leakages of the basement door undercut and transfer grilles. Lastly, results from multizone airflow simulations of the NZERTF, using the calculated leakages as inputs, are presented.

Effective leakages

The ELA at 50 Pa for all 14 tests were calculated using the procedures outlined in ASTM E779 (TABLE 4). The results are listed by test configuration (Test #1 to Test #5) and subconfiguration denoting whether vents and other openings were sealed or unsealed. Because of the attic-outdoor pressure not being able to be neutralized, the results of Tests #3A and #3B actually captured the leakage of the basement wall, first and second floor walls, and the combined leakage of the attic floor and roof. With the attic floor being so leaky relative to the attic roof (see TABLE 5 and subsequent explanation), Test #3 results closely matched the results of Test #1B, which captured the combined leakage of basement wall, first and second floor walls, and attic roof. (The exterior dryer and kitchen exhaust vents were sealed during these three tests.) For Tests #1 and #3, $n = 0.65$ on average, ranging from $n = 0.64$ to $n = 0.67$.

As expected, the ELAs of Tests #2, #4, and #5 (which include the combined leakage of exterior and interior leakages) are greater than the ELAs of Test #1 (exterior envelope only) because attention was paid to minimizing the leakiness of the exterior envelope. No attention was paid to the leakiness of the basement ceiling or attic floor, which is reasonable because both the basement and attic are within the conditioned volume. The ELAs of Test #4 were the highest of the tests, which indicated that the attic floor was leakier than the basement ceiling. This was also verified by further analysis presented below. For Tests #2, #4, and #5, $n = 0.61$ on average (ranging from $n = 0.58$ to $n = 0.68$), which was smaller than the average of the n values for Tests #1 and #3 (tests of the exterior envelope leakage).

TABLE 4. Summary of ELA at 50 Pa for five test configurations and their subconfigurations.

Test number	Leakage determined	ELA at 50 Pa (cm^2)	95 % confidence interval (+/- cm^2)
1A	L_R+L_{BW}+L_W (vents unsealed)	**237**	7
1B	L_R+L_{BW}+L_W (vents sealed)	**200**	3
2A	L_R+L_W + (L_{BW} + L_{BC})[1]	**898**	12
2B	L_R+L_W + (L_{BW} + L_{BC})[1]	**676**	9
2C	L_R+L_W + (L_{BW} + L_{BC})[1]	**765**	8
2D	L_R+L_W + (L_{BW} + L_{BC})[1]	**539**	9
3A	L_{BW}+L_W + (L_R + L_{AF})[1]	**202**	4
3B	L_{BW}+L_W + (L_R + L_{AF})[1]	**204**	3
4A	L_R + L_{AF}[1]	**696**	16
4B	L_R + L_{AF}[1]	**914**	51
4C	L_R + L_{AF}[1]	**694**	17
4D	L_R + L_{AF}[1]	**941**	30
5A	L_{BW} + L_{BC} (first floor grilles sealed)	**611**	4
5B	L_{BW} + L_{BC} (first floor grilles unsealed)	**836**	6

1. See the section, "Test Configurations", for descriptions of the various subconfigurations.

As discussed in the section, "Methodology", the ELAs of the tests are used to calculate C and L of the various building components (i.e., roof, walls, floor). The values of C_R, C_{BW}, and C_W were converted to effective leakages, L, using Eq. (1) and then normalized by their respective surface areas in TABLE 1. The assumption in Eq. (4), $L'_R = L'_W = L'_{BW}$, where the prime notation indicating

L is normalized by surface area, is given in TABLE 5. TABLE 5 also shows the results of calculating L_{AF} and L_{BC} two ways, (1) by determining C_{AF} and C_{BW} and then converting to L and (2) by subtracting the result of the comparable "sealed" test configuration from the "unsealed" test result. TABLE 5 shows the average calculated $L_{\mathrm{AF}} = 618$ cm^2 at 50 Pa (95 % confidence interval (CI) 567 cm^2 to 655 cm^2) when calculated using C_{AF} and shows that the average calculated $L_{\mathrm{AF}} =$ 606 cm^2 at 50 Pa (95 % CI 596 cm^2 to 620 cm^2) when calculated using ELA$_{4C}$ – L_{R}. There was only a 2 % difference in L_{AF} calculated by these two methods.

TABLE 5 shows the average calculated $L_{\mathrm{BC}} = 614$ cm^2 at 50 Pa (95 % CI 594 cm^2 to 634 cm^2) when calculated using C_{BC} and shows the average calculated $L_{\mathrm{BC}} = 611$ cm^2 at 50 Pa (95 % CI 596 cm^2 to 620 cm^2) when calculated using ELA$_{5A}$ – L_{BW} (< 1 % difference). The last column of TABLE 5 shows the effective leakages normalized by their respective surface area or per item. The attic floor (4.66 cm^2/m^2 at 50 Pa) is about 15 % leakier than the basement ceiling (4.01 cm^2/m^2 at 50 Pa) The attic floor and basement ceiling are also about 10 times leakier than the exterior envelope (0.48 cm^2/m^2 at 50 Pa). The leakage area of the attic floor (606 cm^2 at 50 Pa) is greater than the leakage of the attic transfer grilles (233 cm^2 at 50 Pa). The leakage of the basement ceiling (611 cm^2 at 50 Pa) is greater than the leakage of the first floor transfer grilles (224 cm^2 at 50 Pa).

TABLE 5. Flow coefficient and calculated leakages of building components.

	C (m^3/s•Pa)	**L from Eq. (2) (cm^2)**	**L using ELA$_{\mathrm{Test\#}}$ (cm^2)**	**L' (cm^2/m^2 or per item)**
Roof	C_{R}=0.006	L_{R}=88	N/A	0.48 cm^2/m^2
Basement wall	C_{BW}=0.00005	L_{BW}=1		0.48 cm^2/m^2
First and second floor walls	C_{W}=0.010	L_{W}=150		0.48 cm^2/m^2
Attic floor	C_{AF}=0.041	L_{AF}=618	L_{AF}=ELA$_{4C}$ – L_{R} = 606	4.66 cm^2/m^2
Basement ceiling	C_{BC}=0.049	L_{BC}=614	L_{BC}=ELA$_{5A}$ – L_{BW} = 611	4.05 cm^2/m^2

Basement door undercut	N/A	N/A	ELA$_{2A}$ – ELA$_{2C}$ = 132 ELA$_{2B}$ – ELA$_{2D}$ = 137	135 cm^2
First floor transfer grilles (qty = 3)	N/A	N/A	ELA$_{2A}$ – ELA$_{2B}$ = 221 ELA$_{2C}$ – ELA$_{2D}$ = 226 ELA$_{5A}$ – ELA$_{5B}$ = 225	75 cm^2 per transfer grille
Attic transfer grilles (qty = 2)	N/A	N/A	ELA$_{4A}$ – ELA$_{4B}$ = 218 ELA$_{4C}$ – ELA$_{4D}$ = 248	116 cm^2 per transfer grille

SIMULATIONS

To evaluate the effects of different ventilation strategies on airflows and contaminant concentrations in the NZERTF, the *L*' at 50 Pa (cm^2/m^2 or per item) in TABLE 5 were input into a multizone airflow model of the house developed using CONTAM [26]. This CONTAM model was also coupled with EnergyPlus, a whole-building energy analysis tool, to study the energy implications of these airflows [27]. In a previous modeling study, preliminary estimates of the interzone leakage values were used to predict formaldehyde and acetaldehyde concentrations [28].

The CONTAM model considers the interaction between external forces driving airflow (inside-outside temperature difference and wind) and building heating, ventilating, and air conditioning (HVAC) system airflow rates to determine pressures and airflows across internal partitions and the building envelope. CONTAM also accounts for external and internal contaminant sources and removal mechanisms to calculate contaminant transport associated with the airflows. EnergyPlus implements a multizone heat transfer model that accounts for conductive, convective and radiant heat transfer associated with building materials; interzone and envelope airflows; and HVAC systems. During cosimulation using the coupled model, indoor air temperatures and HVAC system airflow rates are passed from EnergyPlus to CONTAM, and airflow rates across the building envelope and between internal zones are passed from CONTAM to EnergyPlus [29, 30]. Details on the model and simulated concentrations of formaldehyde and acetaldehyde in the NZERTF are given in .

Simulations were performed using preliminary estimates for the interzonal leakages that were based on engineering judgement for the floor leakage and manufacturer's catalogs (for the transfer grilles), before the measurements reported on in this paper were performed. The preliminary floor leakage value underestimated the measured value by about half, and the effective leakage value of the transfer grilles had been overestimated by a factor of about three (TABLE 6). Simulations were then repeated using the measured interzonal leakage values. Annual simulations were performed using the Typical Meteorological Year 3 (TMY3) weather file for Baltimore, MD [31], with a time step of 1.0 min and with the heat pump fan controlled by the thermostat (set to 21.1 °C in the heating season and 23.9 °C in the cooling season) and the HRV running continuously at 0.05 m^3/s. Simulations were also performed with both the heat pump fan and HRV off.

TABLE 6. Preliminary and measured leakages of building components.

	Preliminary leakage	***Measured leakage***
Attic floor	2 cm^2/m^2 at 50 Pa	4.66 cm^2/m^2 at 50 Pa
Basement ceiling	2 cm^2/m^2 at 50 Pa	4.05 cm^2/m^2 at 50 Pa
Attic transfer grilles (qty=2)	418 cm^2 /each	116 cm^2/each
First floor transfer grilles (qty=3)	232 cm^2 /each	75 cm^2/each
Basement door undercut	229 cm^2	135 cm^2

TABLE 7 shows that there were significant differences in the predicted airflow rates of the individual paths (averaged over the annual simulation). On average, the predicted flow through the basement ceiling and attic floor using the measured leakage was higher than the value using the preliminary leakage. In contrast, the predicted flow through attic, first floor transfer grilles, and basement door undercut using the measured leakage were lower using the preliminary leakage. Note that the total airflow from the basement to the first floor, and from the second floor to the attic, were the same for the preliminary and measured leakage when averaged over a year (differences < 0.5 %). It may be that no change was observed in the average interzonal airflow because the exterior building leakage was comparatively airtight (ten times more airtight) and changes to the interzonal leakage were not great enough to affect the overall airflow pattern within the house.

The heat pump system was 100 % recirculating and the HRV was balanced. Although there were only heat pump supplies in the basement (no returns), whether the systems were on or off, the total infiltration was the same. With the systems off (details not shown for brevity), the airflow from the basement to the first floor decreased by 4 % because the heat pump was no longer supplying air to the basement. The only outside air supplied to the basement, when the heat pump and HRV was off, would have been through the basement wall, which had an ELA of 1 cm^2 at 50 Pa (0.48 cm^2/m^2 multiplied by the wall area of 2 m^2). With the systems off, the airflow from the second floor to the attic increased by 25 % to balance the decrease of air from the basement to the first floor.

TABLE 7. Predicted interzonal airflow rates using preliminary and measured leakage averaged over a year

Interzonal airflow rates (systems on)	**Average flow using preliminary leakage (m^3/s)**	**Average flow using measured leakage (m^3/s)**	**Percentage difference**
Attic transfer grilles	3.4E-03	9.7E-04	-72 %
Attic floor	1.8E-03	4.2E-03	139 %
First floor transfer grilles	5.3E-05	2.2E-05	-58 %
Basement ceiling	8.2E-06	4.2E-05	415 %
Basement door undercut	3.7E-05	3.4E-05	-9 %
Airflow from second floor to attic	5.2E-03	5.2E-03	-0.11%
Airflow from basement to first floor	9.9E-05	9.9E-05	0.06%

The average simulated concentrations of formaldehyde were not significantly different between using the preliminary and measured interzonal leakage. The differences in the average annual concentrations were < 0.05 % in the basement, combined first and second floor, and attic. This was because the total interzonal airflow between the basement and first floor, and between the second floor and the attic, averaged over a year, did not change whether using the preliminary or measured leakage. The similarity in the simulated formaldehyde concentrations also could have been due to the nature of the entire house being within the conditioned space. In a house with typical-construction, a leaky attic floor (coupled with a vented attic that is not part of the conditioned space) may create a greater stack effect within the house and may redistribute contaminants differently than in the NZERTF. The fact that the interzonal leakage did not affect the distribution of formaldehyde in the NZERTF could have been attributed to the heat pump system recirculating air from the house and delivering it to the basement and the house.

DISCUSSION

One challenge during this series of blower door tests was not being able to neutralize the pressure across the attic roof. Unsuccessful attempts were made to conduct a two-blower test by manually adjusting the speed of the blower door in the front doorway and smaller fan in the attic hatch opening. The authors will attempt a two-blower test using additional fan control devices in the future.

Blower door test results showed that the interior floors were approximately 10 times leakier than the exterior building envelope and that the leakage associated with the transfer grilles between levels was less than the total floor leakage. Considering that all the levels of the NZERTF were all within the conditioned space, the leaky interior floors do not pose an energy penalty (i.e., cold air from the basement flowing up to house, and conditioned air from house flowing out through the roof) when paired with a tight exterior envelope. In a home of more typical construction, however, in which the attic is not within the conditioned space and the exterior walls are not as tight, there could be a significant energy penalty as conditioned air escapes to the attic and out the roof vents.

There were no significant differences in the total interzonal airflows between the levels of the NZERTF, and no significant differences in the simulated formaldehyde concentrations, using preliminary and measured interzonal leakage in the NZERTF (both averaged over a year). Reasons for this lack of difference in the annual averages may include the fact that the zones within the NZERTF were all within the conditioned envelope. Thus cold basements and hot attics did not create as great a stack effect as created in homes with only living areas within the conditioned space. Also, the heat pump in the NZERTF recirculated air between the basement, first floor, and second floor, which may not be the case in all homes. Last, as noted, no change was observed in the average interzonal airflow because the exterior building leakage was comparatively airtight (10 times more airtight) and changes to the interzonal leakage were not great enough to affect the overall airflow pattern within the house. Nevertheless, improved knowledge of interzonal leakage could improve estimates of contaminant transport in other homes, especially when considering transient effects. These findings will be compared with the results of interzonal tracer decay tests, which have been performed in the NZERTF but not yet analyzed.

CONCLUSION

The NZERTF is a 250 m^2 two-story, unoccupied test home located at NIST in Gaithersburg, MD. It is airtight (0.6 h^{-1} at 50 Pa) and has a highly insulated building enclosure designed for heat, air and moisture control. Because the basement, first floor, second floor, and attic levels are all within the conditioned space, no special attention has been given to the airtightness of the interior floors; however, to support airflow modeling efforts, this leakage was quantified through a series of interzonal pressurization tests. It was found that the interior floors were 10 times leakier than the exterior building envelope, and that the leakage associated with the transfer grilles between levels was less than the total floor leakage. This paper described the design of the interzonal tests and the challenges to perform them. Having more accurate estimates of interzonal leakage could be advantageous in understanding the transport of air and contaminants in multizone structures, especially with respect to transient effects.

List of Figure Captions

FIG. 1. NZERTF at NIST facing south.
FIG. 2 Photographs of smoke tests performed at NZERTF at the (a) floor transfer grilles, (b) basement door, and (c) access panel while depressurizing.
FIG. 3. Building surface leakages in NZERTF.
FIG. 4. Five house configurations for determining external and interzonal leakage.
FIG. 5. Electrical circuit equivalent of Test #1
FIG. 6. Electrical circuit equivalent of Test #2.
FIG. 7. Plywood mount for smaller fan in attic hatch opening.

TABLE 1. Physical characteristics of NZERTF

TABLE 2. Summary of subconfigurations for Test #2
TABLE 3. Summary of subconfigurations for Test #4.
TABLE 4. Summary of ELA at 50 Pa for five test configurations and their subconfigurations.
TABLE 5. Flow coefficient and calculated leakages of building components.

REFERENCES

REFERENCES

1. U.S. Energy Information Administration, ed., *Annual Energy Outlook 2018 with Projections to 2050* (Washington, DC: U.S. Energy Information Administration, 2018), 127.
2. *Windows and Building Envelope Research and Development: Roadmap for Emerging Technologies*" (Washington, DC: U.S. Department of Energy, 2014).
3. *Energy Standard for Buildings Except Low-Rise Residential Buildings*, ANSI/ASHRAE/IES Standard 90.1-2016 (Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2016).
4. *U.S. Army Corps of Engineers Air Leakage Test Protocol for Building Envelopes* (Chapaign, IL: U.S. Army Corps of Engineers Research and Development Center, 2012).
5. *International Energy Conservation Code* (Washington, DC: International Code Council, 2015).
6. M. Jayjock and A. A. Havics, "Residential Inter-Zonal Ventilation Rates for Exposure Modeling," *Journal of Occupational and Environmental Hygiene* 15, no. 5 (2018): 376–388, http://doi.org/10.1080/15459624.2018.1438615
7. S. J. Emmerich, J. E. Gorfain, and C. Howard-Reed, "Air and Pollutant Transport from Attached Garages to Residential Living Spaces–Literature Review and Field Tests," *International Journal of Ventilation* 2, no. 3 (2003): 265–276, http://doi.org/10.1080/14733315.2003.11683670
8. L. Du, S. Batterman, C. Godwin, J.-Y. Chin, E. Parker, M. Breen, W. Brakefield, T. Robins, and T. Lewis, "Air Change Rates and Interzonal Flows in Residences, and the Need for Multi-Zone Models for Exposure and Health Analyses," *International Journal of Environmental Research and Public Health* 9, no. 12 (2012): 4639–4662, http://doi.org/10.3390/ijerph9124639
9. E. L. Hult, M. H. Sherman, I. S. and Walker, "Blower-Door Techniques for Measuring Interzonal Leakage" (paper presentation, *Buildings XII*, Lawrence Berkeley National Laboratory, Berkeley, CA, December 4-6, 2013).
10. C. F. A. Afonso, E. A. B. Maldonado, and E. Skåret, "A Single Tracer-Gas Method to Characterize Multi-Room Air Exchanges," *Energ Buildings* 9, no. 4 (1986): 273–280.
11. F. D. Heidt, R. Rabenstein, and G. Schepers, "Comparison of Tracer Gas Methods for Measuring Airflows in Two-zone Buildings," *Indoor Air* 1, no. 3 (1991): 297–309.
12. S. L. Miller, K. Leiserson, and W. W. Nazaroff, "Nonlinear Least-Squares Minimization Applied to Tracer Gas Decay for Determining Airflow Rates in a Two-Zone Building," *Indoor Air* 7, no. 1 (1997): 64–75, http://doi.org/doi:10.1111/j.1600-0668.1997.t01-1-00008.x

13. P. J. O'Neill and R. R. Crawford, "Identification of Flow and Volume Parameters in Multi-Zone Systems Using a Single-Gas Tracer Technique," *ASHRAE Transactions* 1991, no. 1 (1997): 49–54.
14. M. Enai, N. Aratani, C. Y. Shaw, and J. T. Reardon (1993). Differential and integral method for computing interzonal airflows using multiple tracer gases. *Proceedings of International Symposium on Room Air Convection and Ventilation Effectiveness (ISRACVE), 22-24 July, 1992 Tokyo, Japan, Atlanta: ASHRAE*, 357-362.
15. R. P. Sieber, R. W. Besant, and G. J. Schoenau, "Variations in Interzonal Airflow Rates in a Detached House Using Tracer Gas Techniques" r presentation, *ASHRAE Transactions* 99, (1993): 699–708.
16. L. Du, S. Batterman, C. Godwin, Z. Rowe, and J. Y. Chin, "Air Exchange Rates and Migration of VOCs in Basements and Residences," *Indoor Air* 25, no. 6 (2015): 598–609, http://doi.org/10.1111/ina.12178
17. A. H. Fanney, V. Payne, T. Ullah, L. Ng, M. Boyd, F. Omar, M. Davis, H. Skye, B. Dougherty, B. Polidoro, W. Healy, J. Kneifel, and B. Pettit, "Net-Zero and Beyond! Design and Performance of NIST's Net-Zero Energy Residential Test Facility," *Energy and Buildings* 101, no. 15 (2015): 95–109, http://doi.org/10.1016/j.enbuild.2015.05.002
18. F. Omar and S. T. Bushby, "Simulating Occupancy in the NIST Net-Zero Energy Residential Test Facility" (Gaithersburg, MD: National Institute of Standards and Technology, 2013).
19. L. Ng, A. Persily, and S. Emmerich (2015), “Infiltration and Ventilation in a Very Tight, High Performance Home” *Proceedings of 36th AIVC Conference Effective Ventilation High Performance Buildings*, Madrid, Spain, September 23–24, 2015, Brussels: AIVC, 719–726.
20. *LEED BD+C: Homes, v4*, LEED v4 (Washington, DC: U.S. Green Building Council, 2014).
21. *ENERGY STAR Certified Homes, Version 3.1* (Rev. 06) (Washington, DC: U.S. Environmental Protection Agency, 2015).
22. *PHIUS + 2015: Passive Building Standard–North America*" (Chicago, IL: Passive House Institute US, 2015).
23. W. R. Chan, J. Joh, and M. H. Sherman, "Analysis of Air Leakage Measurements of US Houses," *Energy and Buildings* 66 (November 2013): 616–625, http://doi.org/10.1016/j.enbuild.2013.07.047
24. *ASHRAE Handbook Fundamentals* (Atlanta, GA: American Society of Heating, Refrigerating and Air-Conditioning Engineers, 2017).
25. *Standard Test Method for Determining Air Leakage Rate by Fan Pressurization* ASTM E779-19 (West Conshohocken, PA: ASTM International, 2019), www.astm.org
25. B. Pettit, C. Gates, A. H. Fanney, and W. Healy, "Design Challenges of the NIST Net-Zero Energy Residential Test Facility" (Gaithersburg, MD: National Institute of Standards and Technology, 2014)..
26. W. S. Dols and B. Polidoro, "CONTAM User Guide and Program Documentation" (Gaithersburg, MD: National Institute of Standards and Technology, 2016).
27. *EnergyPlus 8.4* (Washington, DC: U.S. Department of Energy, 2015).
28. L. Ng, D. Poppendieck, W. S. Dols, and S. J. Emmerich, "Evaluating Indoor Air Quality and Energy Impacts of Ventilation in a Net-Zero Energy House Using a Coupled Model,"

Science and Technology for the Built Environment 24, no. 2 (2017): 124–134, https://doi.org/10.1080/23744731.2017.1401403

29. W. S. Dols, S. J. Emmerich, and B. J. Polidoro, "Coupling the Multizone Airflow and Contaminant Transport Software CONTAM with EnergyPlus Using Co-simulation," *Building Simulation* 9 (2016): 469–479, http://doi.org/10.1007/s12273-016-0279-2
30. M. Wetter, "Co-simulation of Building Energy and Control Systems with the Building Controls Virtual Test Bed," *Journal of Building Performance Simulation* 4, no. 3 (2011): 185–203, http://doi.org/10.1080/19401493.2010.518631
31. "National Solar Radiation Data Base: 1991–2005 Update: Typical Meteorological Year 3," (Lakewood, CO: National Renewable Energy Laboratory, 2015).

An ASABE Meeting Presentation

DOI: https://doi.org/10.13031/aim.201801319

Paper Number: 1801319

Implementing Grain Traceability Standards: CART and Simulation

Riddick, Frank H.[1]; Wallace, Evan K.[1]; Nieman, Scott;[2] Tevis, Joe W.[3]; Ferreyra, R. Andres[4,*]

[1] National Institute of Standards and Technology, Gaithersburg, MD; [2] Land O'Lakes, Inc, Shoreview, MN;
[3] Vis Consulting, Inc., Waconia, MN; [4] Ag Connections, LLC, Murray, KY
**** Corresponding Author***

Written for presentation at the
2018 ASABE Annual International Meeting
Sponsored by ASABE
Detroit, Michigan
July 29-August 1, 2018

ABSTRACT. ***To ensure food safety, food manufacturers need the ability to quickly identify and trace food products to all equipment and processes throughout the entire associated food supply, production, and transportation network. Government and industry have recognized that supply chain traceability is the key to keeping the public safe, but new information standards are needed to enable such traceability. This paper describes several efforts being undertaken by industry, government, and academic organizations to develop and standardize traceability technology, focusing on grain traceability, and how modeling, simulation, and analysis technology are being used to support these projects. Noteworthy among these efforts is AgGateway's Commodity Automation by Rail and Truck (CART) project, which sought to understand the business processes and data exchanges required, all the way from farm operations through grain elevators to receiving at a feed manufacturer; exploring solutions through small proof-of-concept (PoC) projects; and enhancing and implementing existing standards, primarily the ISO 11783 standard for farm machinery electronics, and the AgXML standard for grain data exchange. The paper also presents ongoing work on using agent-based simulation to validate proposed traceability standards.***

Keywords. agent-based simulation, food traceability, standardization, supply chain analysis.

Introduction

Negative impacts to the food supply chains can be extremely costly, having caused over $22 billion in losses due to food recalls and related expenses (Hussain and Dawson, 2013). They can also be dangerous, having caused over 8.9 million illnesses, 53,245 hospitalizations, and over 2300 deaths (Flynn, 2014; ERS, 2017). Managing the supply chain for complex products such as cars or airplanes is difficult and might involve millions of parts from dozens of suppliers. Managing the supply/production chain in the food manufacturing domain adds a unique characteristic of both the input materials and the final product: they are biologically-based and thus perishable; they are usable during a limited timespan. This requires special procedures for gathering, storing, handling, and transporting materials and end-products, making the logistics for input sourcing and end-product distribution complex, costly, and susceptible to a myriad of disturbances including adverse weather

Riddick, Frank H.; Wallace, Evan K.; Nieman, Scott; Tevis, Joe; Ferreyra, R. "Implementing Grain Traceability Standards: CART and Simulation." Paper presented at 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Detroit, MI, US. July 29, 2018 - August 01, 2018.

conditions and transportation worker strikes.

There is constant variability in the characteristics of food production input materials. In addition, vast differences can be observed in the material characteristics of a product across lots, even from the same year. This forces manufacturers to constantly test, monitor, and find alternate sources for their input materials and then modify their products' manufacturing processes based on the characteristics of the materials that they were able to source.

To help manage the complexity of the food supply production chain and to deal with adverse events when they occur, most participants in the supply production chain have traceability-enabling systems. Briefly, traceability is "the ability to track any food through all stages of production, processing and distribution" (European Commission, 2017). Unfortunately, there is no standard for traceability systems or their information content that covers the complete food production chain. Furthermore, at many points in the supply chain, traceability information is collected and recorded manually and maintained on physical records, severely limiting the speed with which the information can be accessed when needed.

A system providing end-to-end traceability would enable the discovery and use of synergies in the food production chain; this would lead to a more efficient and cost-effective supply and production system; and would enable that problems be addressed quickly and with a minimum effort. The need for such a system has been recognized; many companies work to improve traceability within their organizations (Church, 2015; Dow, 2018) and academic researchers are studying various aspects of the traceability problem (Thakur and Hurburgh, 2009). Through the passage of the Food Safety Modernization Act (FSMA; US FDA, 2018) the U.S. government has mandated basic procedures that stakeholders in the food production and supply chain must follow, but because a standard system to support traceability does not exist, dates for strict compliance with FSMA have been delayed (Schultz, 2018).

To move towards the goal of a standards-based system supporting end-to-end traceability, several organizations have efforts underway to address traceability related deficiencies in the current food supply production chain. This paper describes several of these efforts, emphasizing how modeling, simulation, and analysis technology can help reach the desired goals. The technical developments needed to support end-to-end traceability are also discussed, as well as how the results of current traceability research efforts will be coordinated to support future efforts.

Key Notions of Traceability

The UN definition of traceability, "The ability to identify and trace the history, distribution, location and application of products, parts and materials…", points to a key function of traceability systems: *trace back* (UN, 2014). Trace back is the ability to assemble, for a given product, information about the input materials, processes, and human or machine resources that were directly used to create the product, and then to recursively repeat that process for each of the materials, processes, and/or human and machine resources discovered. This process is extremely important when trying to uncover the time, location, and nature of the adverse event in the production chain that caused a product defect.

A complement to the trace back function of traceability systems is the ability to *trace forward*. This function seeks to, given a specific material and/or piece of equipment at a specific point in the production chain, identify the materials at the next step in the production chain that used the previous materials as inputs or used the associated equipment in its processing, repeating this process recursively until all end products that are directly or indirectly related to the initial material or equipment have been identified. The trace forward function is necessary to identify the scope of potential end-products affected once the root cause and origin of a product defect has been determined.

One problem that makes traceability difficult is that different partners in the supply/production chain might use different management and/or computer systems to carry out their operations, and these systems might use different information to identify or describe the same product or material. Some kinds of traceability analyses only look at information describing how products move between partners; others focus on each traceability-related activity that happens within a partner's organizational boundary. Issues, information, and analysis focused on the inter-partner production chain activities are referred to as being a part of *external traceability* while issues, information, and analysis focused on what happens within a partner's organizational boundary are referred to as being a part of *internal traceability*.

The information necessary to support a traceability system is complex, interrelated, and vast; making sense of it can be very challenging. Even though a clear goal of a traceability system is to be able to associate a product with its inputs, the current approach most widely used to organize traceability information focuses on the events that occur within a production chain that will enable product inputs and outputs to be traced. With this approach, *Critical Tracking Events* (or CTEs) are defined to contain data about the input and output materials, processes, locations, and human or machine resources associated with each event in the production chain that is necessary to support the trace forward and trace back functions of a traceability system (Badia-Melis et al., 2015).

Using CTEs as the basis for recording, managing, and analyzing production chain data has been shown to enable supporting traceability analyses related to external traceability, internal traceability, and analyses which include both internal

Riddick, Frank H.; Wallace, Evan K.; Nieman, Scott; Tevis, Joe; Ferreyra, R. "Implementing Grain Traceability Standards: CART and Simulation." Paper presented at 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Detroit, MI, US. July 29, 2018 - August 01, 2018.

and external elements.

Barriers to the Development of Traceability Systems for Bulk Grain

The continued occurrence of food safety incidents has only underscored the need for traceability systems that can quickly perform trace back and trace forward functions to limit the exposure of consumers to serious illnesses and possible death. Developing and deploying such systems present a variety of problems. Developing systems to support production chains that have bulk grain as one of its materials add to those problems; some specific challenges are detailed below.

Difficulties in conceptualizing an end-to-end traceability system

Since no end-to-end traceability system currently exists, it is difficult to even discuss the concept of such a system with many production chain partners. Partners may have only focused on the internal traceability issues of their organizations and much of that information may only be maintained as tribal knowledge within the organization. Also, efforts to support external traceability may be limited to providing “one-up, one-down” traceability for themselves and their direct production chain partners (Bhatt et al., 2013).

No standard promoting consistency in the representation and interpretation of traceability information

In general, each participant in a production chain has their own unique enterprise and operational infrastructure. Traceability system components will need to be built upon the different infrastructural components of each partner, and mechanisms are needed to provide consistency and reduce or eliminate ambiguity in the information created and exchanged within the production chain. First, there is a need to standardize how to represent the data describing traceability concepts like events, materials, resources, and processes. Second, there is a need to provide context given that the same term may be interpreted differently by different production chain participants. This becomes evident by observing that the term “lot” could refer to a bag of grain, a truckload of grain, a rail car full of grain, of even a portion of land on which grain is grown.

Lack of methods for assessing relationships between input and output grain lots passing through grain storage bins

In production chains involving grains such as wheat or corn, the grain is often stored in large containers referred to as *bins* or *silos*. In general, these bins operate like queues: grain enters through the top and comes out of the bottom. Unfortunately, due to many factors (e.g., bin geometry, type of grain, grain moisture content, and extraction method) problems such as doming and rat-holing cause the grain movement through the bin to diverge from a strict first-in, first-out material movement pattern. In order to assess which bin outputs are connected with specific bin inputs, methods need to be developed to characterize how movement through a bin is affected by bin geometry and/or bin environmental conditions.

No methods for evaluating proposed solutions for defining, exchanging, and analyzing traceability information

As candidate solutions for the aforementioned problems are developed, methods must be developed to assess whether traceability system based on these solutions would operate acceptably before the solutions are deployed.

Addressing Traceability Needs for Production Chains Involving Bulk Grains

The need for systems, methods, and standards to enable the deployment of traceability systems has been recognized by government, academic organizations, and the stakeholders in the food production chain, both in the U.S. and abroad (Donnelly and Thakur, 2010; Thakur et al., 2009; Opara, 2003). The motivation for feasible, economical, standards-based solutions for traceability has recently increased due to the lack of solutions that stakeholders in the U.S food production chain can use to comply with the precepts of FSMA. Several academic, government, and industry organizations, which were individually working on different aspects of the problem for production chains involving bulk grains, have recently begun to work together to develop methods and information specifications for traceability systems that might lead to standardized traceability solutions. While this is not an official “project”, these organizations hope that collaborating will result in synergy and lead to deployable results for each of their efforts. Several of these efforts are described below.

Modeling processes and their data: BPMN and Transfer Event modeling

Enabling all stakeholders to understand the complexities of an extensive production chain is a difficult but necessary task. Often, information about a process early in the chain must be collected and maintained to enable the ability to trace back from a downstream process. To illustrate how each production chain stakeholder’s processes interact, participants in the AgGateway industry consortium’s interoperability projects have been using Business Process Model and Notation (BPMN; Silver, 2011) to model the production chain (AgGateway, 2017). They were joined in this effort by researchers from the Open Applications Group Quality Content work group and from the NIST Agri-food Manufacturing System and Supply Chain Integration project, who have been doing similar modeling for supply and production chains in other domains. This collaboration involved modeling the overall production chain and key processes along it (OAGi, 2016; NIST 2017). As an example, Figure 1 presents a BPMN diagram that illustrates the harvesting process for bulk grain, created using a web-based BPMN modeling tool (Trisotech, 2018).

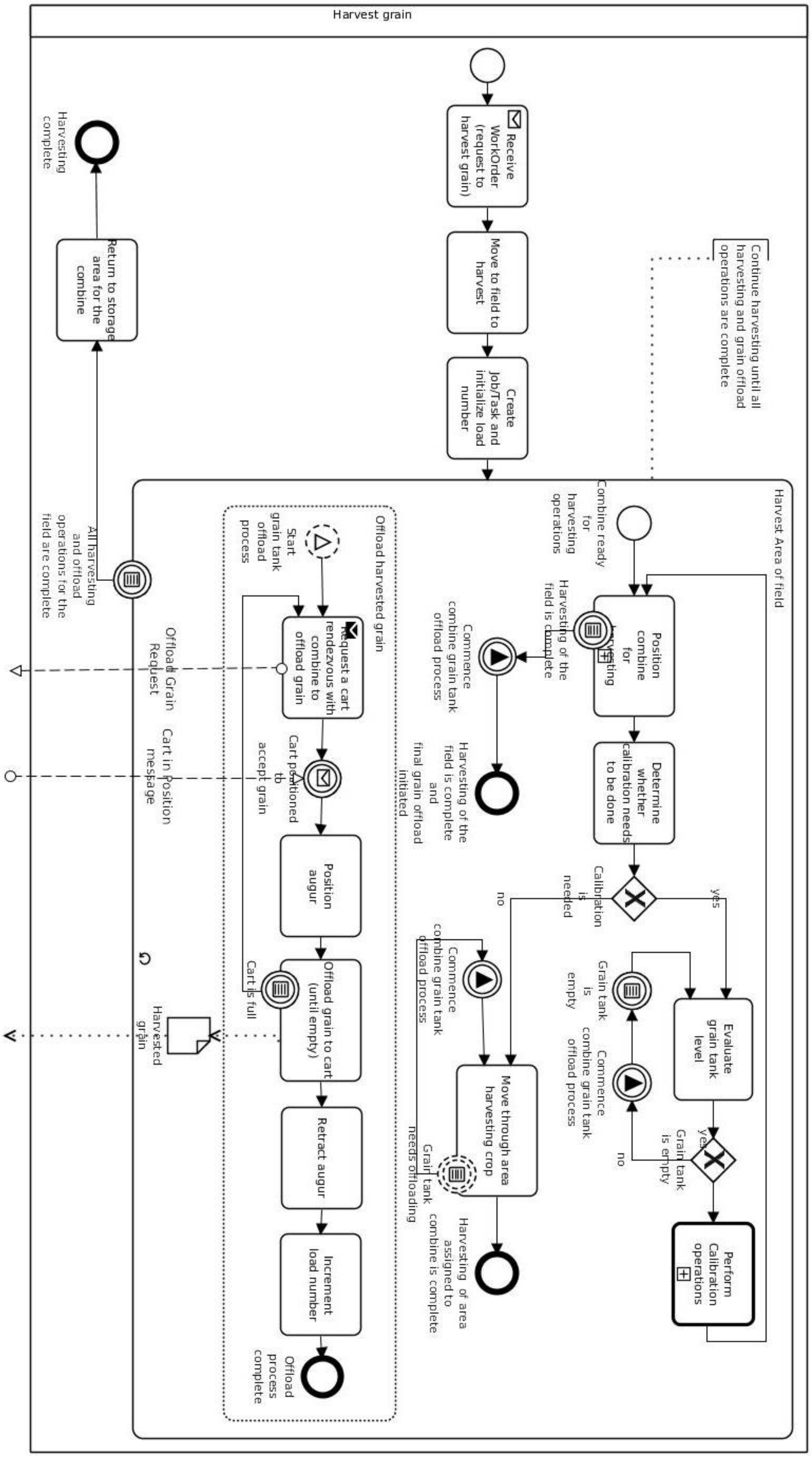

Harvest grain
Receive WorkOrder (request to harvest grain)
Move to field to harvest
Create Job/Task and initialize load number
Continue harvesting until all harvesting and grain offload operations are complete
Harvest Area of field
Combine ready for harvesting operations
Position combine for harvesting
Harvesting of the field is complete
Commence combine grain tank offload process
Harvesting of the field is complete and final grain offload initiated
Determine whether calibration needs to be done
Calibration is needed
yes
no
Evaluate grain tank level
Grain tank is empty
Commence combine grain tank offload process
yes
no
Perform Calibration operations
Commence combine grain tank offload process
Move through area harvesting crop
Grain tank needs offloading
Harvesting of area assigned to combine is complete
Offload harvested grain
Start grain tank offload process
Request a cart rendezvous with combine to offload grain
Offload Grain Request
Cart positioned to accept grain
Cart in position message
Position augur
Offload grain to cart (until empty)
Cart is full
Harvested grain
Retract augur
Increment load number
Offload process complete
All harvesting and offload operations for the field are complete
Return to storage area for the combine
Harvesting complete

Figure 1: BPMN process model of a grain harvest, including subprocesses for harvesting an area of the field and

offloading grain to a cart.

The data elements that describe the depicted processes were established and documented using BPMN diagrams of key production chain processes as a guide. With respect to support for traceability, it was decided that the focus should be on the critical tracking events and supporting data elements that document the transfer of quantities of grain from one container to another. Figure 2 shows how such a CTE named *TransferEvent* could be represented as an extension of the AgXML standard (AgXML, 2009) which defines information entities that can be used to exchange data between grain production chain partners.

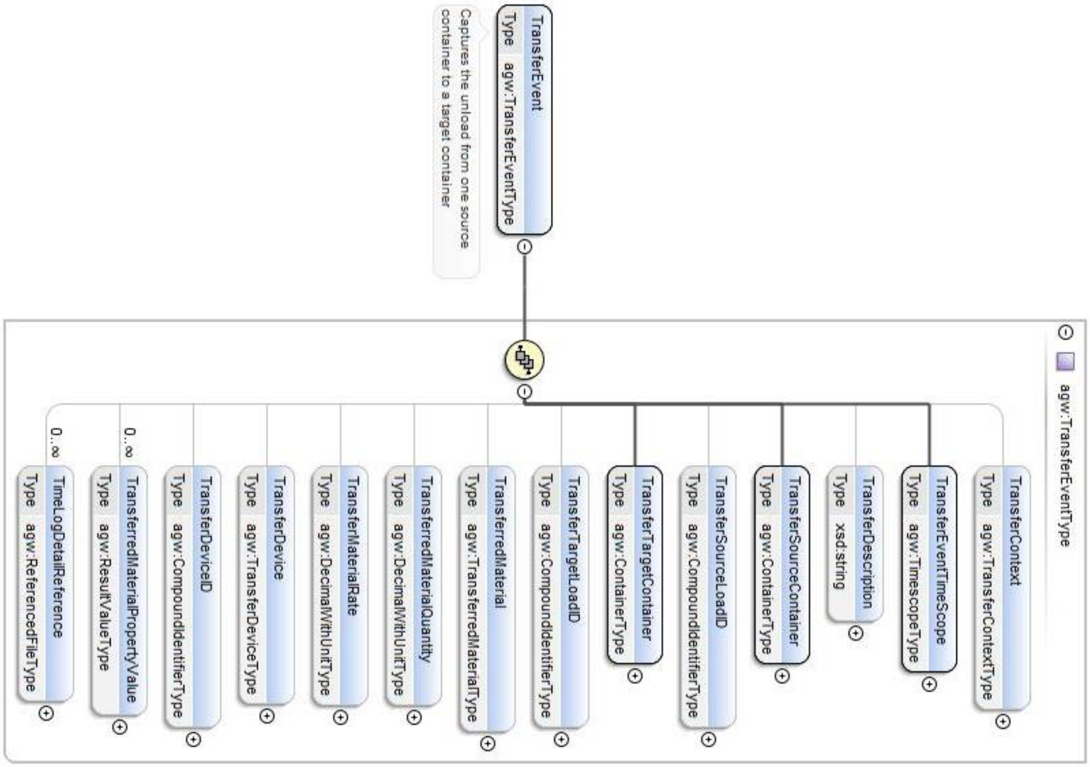

TransferEvent
Type agw:TransferEventType
Captures the unload from one source container to a target container
agw:TransferEventType
0..∞
TimeLogDetailReference
Type agw:ReferencedFileType
0..∞
TransferredMaterialPropertyValue
Type agw:ResultValueType
TransferDeviceID
Type agw:CompoundIdentifierType
TransferDevice
Type agw:TransferDeviceType
TransferMaterialRate
Type agw:DecimalWithUnitType
TransferredMaterialQuantity
Type agw:DecimalWithUnitType
TransferredMaterial
Type agw:TransferredMaterialType
TransferTargetLoadID
Type agw:CompoundIdentifierType
TransferTargetContainer
Type agw:ContainerType
TransferSourceLoadID
Type agw:CompoundIdentifierType
TransferSourceContainer
Type agw:ContainerType
TransferDescription
Type xsd:string
TransferEventTimeScope
Type agw:TimescopeType
TransferContext
Type agw:TransferContextType

Figure 2: TransferEvent as an extension to AgXML

AgGateway Commodity Automation for Rail & Truck (CART) Project Proof of Concept (POC)

To overcome deficiencies in the hardware, software, and data storage and exchange systems available to support production chain traceability and information exchange automation, AgGateway, an industry consortium with 200+ members, dedicated to enabling digital agriculture, started a project called the Commodity Automation for Rail & Truck (CART) project. The stated goal of CART is to facilitate “grain traceability from combine to grain cart, to truck, to elevator, to food processor” (AgGateway, 2018). To reach that goal, CART is conducting a number of technology, or proof of concept (POC) demonstrations, to evaluate the feasibility of proposed solutions for traceability. These demonstrations are in fact live simulations as defined by the live, virtual, constructive simulation taxonomy (US DoD, 1998). The POCs involve both modified and unmodified hardware and software, integrated together to collect, store, and exchange traceability information collected during the execution of real grain harvest events. Participants in CART include: farm operators, hardware and software vendors (some of which modified their products to participate in the POC), grain elevator operators, and processors (manufacturers of end-products that use grain as an input). The key processes of the harvest event were illustrated by diagrams such as Figure 1, and the transfer event data was exchanged as specified by the representation depicted in Figure 2. The points in the production chain where TransferEvents are exchanged are depicted in Figure 3.

The POC enabled participants to evaluate many different aspects of the proposed traceability solution, including overall approach feasibility, existing sensor adequacy for raw data collection, proposed hardware and software modification feasibility, and cloud-based solution feasibility for data storage and exchange. Some of the results of the initial POC are:

- The overall approach is feasible;
- TransferEvent data should be adequate to support traceability needs, and;

Riddick, Frank H.; Wallace, Evan K.; Nieman, Scott; Tevis, Joe; Ferreyra, R. ”Implementing Grain Traceability Standards: CART and Simulation.” Paper presented at 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Detroit, MI, US. July 29, 2018 - August 01, 2018.

- Hardware may need to be enhanced with additional automation capabilities and additional hardening to withstand environmental conditions.

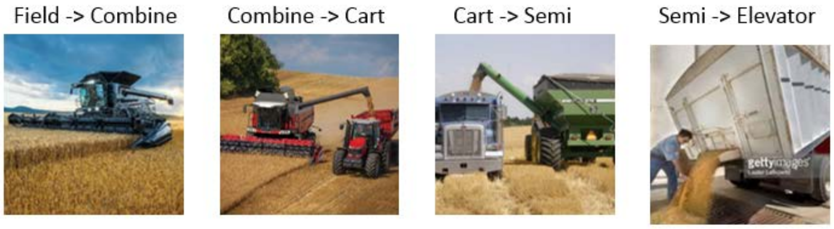


Figure 3: Points in the process where TransferEvents are exchanged

More information about the initial POC can be found in AgGateway (2017A). An additional POC is planned for Fall 2018.

Traceability Through Facilities Typified as Multi-Storage-Bin Environments

Facilities such as grain elevators typically have multiple storage bins to hold grain until dispensed for downstream partners in the production chain. With respect to traceability, two challenges commonly occur due to standard operating procedures at such facilities. First, delivered grain may be distributed to multiple bins and typically records documenting which bins received grain and the amount of grain received may not be kept permanently and in a searchable form. Second, there are no methods for tying grain inputs to outputs based on expected bin flow conditions, taking into account environmental (temperature, humidity, etc.), material (grain type, particle size, moisture content, etc.) and bin shape, material, etc.

To address these issues, NIST is funding research at Iowa State University to look into the traceability issues affecting bulk grains. Part of that effort involves analysis of past and current efforts in defining and using CTEs as the basis for supporting traceability. The output of this research has provided input to the efforts for CTE modeling and standardization described in the previous sections. In addition, part of this effort involves analyzing methods for bin flow characterization towards the goal of developing a mathematical model that can tie bin input to outputs. These efforts are in the early stages, the goal is to be able to use the output of this research to enhance capabilities of tools that support bulk grain traceability.

Simulating Production Chain Traceability for Illustration and Analysis

As previously mentioned, current research suggests traceability in production chains would be best supported by defining and maintaining CTE data about the important events that took place due to production chain operations. As with the live simulation events described previously, efforts to verify the feasibility of proposed solutions for traceability need to take place before standardization and deployment of those solutions. To support this type of analysis, data sets need to be constructed containing realistic collections of CTEs covering the operations of all stakeholders in the production chain and adhering to the format and content proposed by those solutions. To verify that a solution could support traceability analysis, CTE data associated with a specific production chain operation could be modified to indicate the occurrence of an adverse event, and then propagated through the production chain until a set of affected end products is identified. Then, trace back and trace forward analysis could be attempted to determine if all affected products, with few false positives, could be identified from the CTE data defined according to the proposed solution.

To be able to evaluate different proposed content specifications for CTE data, researchers at NIST have created a simulation of a bulk grain production chain that can generate CTE data sets. The primary goal is to generate a realistic collection of TransferEvent data, covering all transfers of grain from the field to the processor, to determine the extent to which traceability analysis can identify end-products affected by adverse events in the production chain. Secondary goals are to illustrate how individual stakeholders interact to perform production chain operations, and to provide a means to evaluate how different models of grain flows through storage bins might affect a proposed solution's traceability capabilities.

The simulated scenario is the same as that covered by the POC live simulation presented above: combines harvest grain from a field; when full, each combine offloads its grain to a cart; when a cart is full, it offloads to a truck; when full, a truck will deliver to an elevator where the grain is transferred to a storage bin; when a processor needs grain, it will send a truck to an elevator where grain is offloaded from a storage bin to the truck; this will then deliver to the processor, where grain will be offloaded to the processor's storage bin. Each time grain is moved from one container to another, a TransferEvent record should be created. In this scenario, not only will grain movement between obvious containers generate a TransferEvent (e.g., grain transfer from a combine's storage compartment to a cart's storage compartment) but a combine's harvesting of grain will also generate a TransferEvent – the area of a field from which the grain is harvested is considered a container for traceability purposes. This matches the TransferEvent generation points described in Figure 3. As was the case in the POC, the content for TransferEvents is based on the proposed specification described in Figure 2, although in both cases data for most of the optional fields was not generated. The behavior for the part of the simulation that covers operations from the field up to the loading of the semi-truck adheres to that specified by the diagram in Figure 1.

In addition to gathering information directly from participants in the POC and others involved in production chain operations, preparation for creating the simulation involved analyzing equipment specifications to determine appropriate value ranges for the key equipment characteristics that would be needed to simulate production chain operations. Information about typical farm crop characteristics was gathered from www.usda.gov.

The simulation was created using the AnyLogic software program (AnyLogic, 2018). This software enables the creation of multimethod simulations; the simulation described here was created as a hybrid agent-based and discrete-event simulation. Agents are constructed to represent the behavior of key entities in the scenario being simulated. The current simulation covers harvesting operations from the field to the dispatching of a Truck to an elevator. Agents to cover operation at the elevator and at the processor are currently under development. Details about several of the currently implemented agents can be found in Table 1.

Table 1: Agents and their responsibilities

Agent	Responsibilities	Key attributes
FieldMgr	Based on initialization data: • Create **Combine, Cart, Truck,** and **CropZone** agents • Assign/reassign **Combines** to **CropZones** • Manage how **Trucks** are filled with grain from **Carts** • Provide visualization of the state of harvesting operations	name id myFarm myGrower
CropZone	• A part of a field that is to be harvested by a **Combine** • CropZone is further subdivided into Areas • Areas enable crop conditions (e.g., moisture, quality, or yield potential) to be varied at a sub-**CropZone** level • Area size can be varied depending on the fidelity need for traceability analysis at the field level • Provide a specific location in a **CropZone** where an adverse field condition can be introduced	name id myFarm myGrower myField AreaList: List [0..n] of Area assignedCombine
Combine	• Seize a **CropZone** • Iteratively process each area in a **CropZone** until finished • The time to finish harvesting each area and the amount of grain harvested are determined stochastically • Track the amount of grain harvested • Pause harvesting and request a **Cart** for offloading operations when the **Combine's** hopper is full • Collaborate with **Cart** and **CartHandler** to perform offloading operations • Resume harvesting once the hopper is empty • Request a new **CropZone** to process after finishing the current one if more **CropZones** need harvesting	myFarm myField zoneTransferSpeed (mph): (time necessary to go from one zone to another.) harvestSpeed (bu/sec) movementSpeed (mph) hopperCapacity (bu) curHopperAmount(bu) areaGrainYield (bu/acre) offloadSpeed (bu/sec)
Cart	• Accept grain from **Combines** and, once full, offload it to a **Truck** • Track current grain amount in its storage compartment • Collaborate with **Combine** and **CartHandler** to perform offloading operations from the **Combine** • Collaborate with **Truck** and **CartHandler** to perform offloading operations to the **Truck**	curZone unloadSpeed (bu/sec) movementSpeed (mph) curAmount(bu) curPartner moveToCombineTime MoveToTruckTime
CartHandler	• Accept grain from **Combines** and, once full, offload to a **Truck** • Collaborate with **Combine** and **Cart** to perform offloading operations from the **Combine** • Collaborate with **Truck** and **Cart** to perform offloading operations to the **Truck**	curZone unloadSpeed (bu/sec) curCart curCombine curTruck

Truck	• Collaborate with **Cart** and **CartHandler** to perform offloading operations • Once full, travel to a **Processor** to offload its grain • Return to original location after grain delivery is complete	Name id movementSpeed (mph) curAmount(bu) location

The execution of the simulation takes place by agents carrying out their defined behaviors and through agent interaction. Agent behavior may be defined by a network of interconnected process modeling blocks (e.g., source, sink, or delay). This is typical for discrete-event simulation software. In addition, agent behavior can be defined by timed, message-based, or condition-based transitions between states in a state chart defined for the agent. In this case, specific behaviors are defined in functions written in the Java language-based coding framework of the AnyLogic software package. These functions are executed based on transitions firing or with entering or exiting a state. Figure 4 shows the state chart associated with a Truck agent.

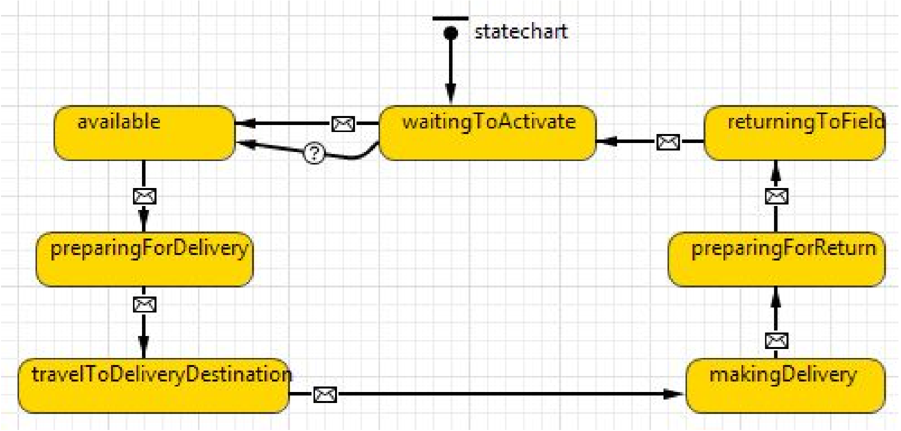


Figure 4: Chart showing the various states for a Truck agent in the simulation. Transitions between states are driven by messages. The truck is initially in a waiting to activate state and then transitions sequentially through the steps required to pick up material to the farm, travel to an elevator, deliver at the elevator and return to the original state to start the cycle again.

This simulation is not intended to provide a realistic emulation of the actual movement of individual vehicles in a harvesting operation. It is based on stochastic variation for the key events, including TransferEvents, that take place during harvesting and operation of the rest of the production chain. This approach: (1) provides a means to produce TransferEvent data sets that enable traceability analysis of the grain movement events in a complex production chain with many participants, and; (2) makes it easier to create simulations of different production chains by just changing the number and initialization data for the agents.

Figure 5 provides a snapshot of a simulation execution with two combines, two carts, three trucks, and four crop zones. The logic that governs a part of the behavior of Combine and Cart agents is presented at the top of the figure. Illustrations for the current states for the Combine, Cart, and Truck agents are provided below that. On the left, the current status for harvesting operations on each CropZone is provided. CropZones for which harvesting has not started, is underway, or has completed are color-coded in light green, medium green, or yellow, respectively. Note that since only two Combines are available, some will have to harvest more than one CropZone. In this case, the combine *Combine_1* was assigned to and finished harvesting cropzone *Zone_2* and is currently harvesting cropzone *Zone_4*.

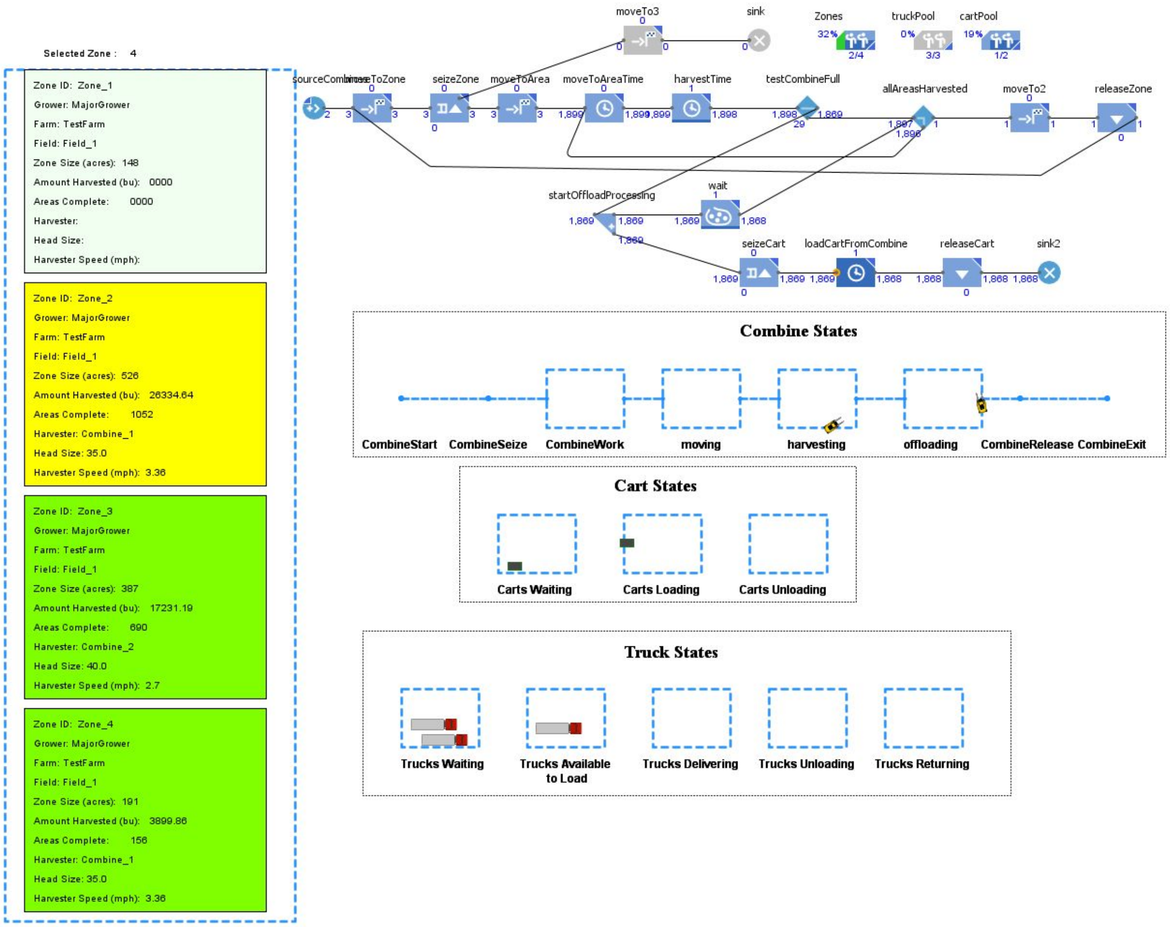
Summary and Future Work

In this paper, the nature and goals of several projects that are addressing traceability issues have been described. The projects seek to gain synergy through inter-project collaborations, which should both increase the quality and confidence in the outputs of the individual projects. The participants plan to continue to collaborate on areas of common interest. Future activities may include: identification of new stakeholders that can contribute to the different efforts; continued evaluation of the data needs to support traceability; identification and evaluation of existing standards to assess how they might be used to construct systems that support traceability; exploration of whether the development of a cross domain ontology for traceability is feasible, and; finishing work on the modeling of grain flow through a bin.

Analysis of the results of the POC continue, and at least one additional live simulation of the grain production chain is planned, with many new participants showing interest. The target for this effort is Fall 2018.

Riddick, Frank H.; Wallace, Evan K.; Nieman, Scott; Tevis, Joe; Ferreyra, R. ”Implementing Grain Traceability Standards: CART and Simulation.” Paper presented at 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Detroit, MI, US. July 29, 2018 - August 01, 2018.

Immediate plans are to extend the production chain simulation to cover all production chain participants from the farm to the processor. This will enable it to generate a data set of transfer events that can be analyzed that covers the same parts of the production chain as covered by the POC. Future extensions of the simulation include: reading initialization data from and storing results in a database to make it easier to analyze the data; creating a form based front end for the creation of initialization data to reduce the effort to create different production chain scenarios to simulate, and; integrating the result of the bin flow modeling effort once they are available.



Selected Zone : 4

Zone ID: Zone_1
Grower: MajorGrower
Farm: TestFarm
Field: Field_1
Zone Size (acres): 148
Amount Harvested (bu): 0000
Areas Complete: 0000
Harvester:
Head Size:
Harvester Speed (mph):

Zone ID: Zone_2
Grower: MajorGrower
Farm: TestFarm
Field: Field_1
Zone Size (acres): 526
Amount Harvested (bu): 26334.64
Areas Complete: 1052
Harvester: Combine_1
Head Size: 35.0
Harvester Speed (mph): 3.36

Zone ID: Zone_3
Grower: MajorGrower
Farm: TestFarm
Field: Field_1
Zone Size (acres): 387
Amount Harvested (bu): 17231.19
Areas Complete: 690
Harvester: Combine_2
Head Size: 40.0
Harvester Speed (mph): 2.7

Zone ID: Zone_4
Grower: MajorGrower
Farm: TestFarm
Field: Field_1
Zone Size (acres): 191
Amount Harvested (bu): 3899.86
Areas Complete: 156
Harvester: Combine_1
Head Size: 35.0
Harvester Speed (mph): 3.36

Figure 5: Snapshot of graphics that illustrate the simulation including (going clockwise, starting from top right): a graph illustrating a portion of the logic governing the behavior of the Combine and Cart agents; (below that) three shapes illustrating major states for Combine agents, Cart agents, and Truck agents respectively, with icons indicating the agents in the simulation in those states; and (on the left side) CropZones and the current states of those zones in the simulation with different colors: white [or light green] for harvesting not started, yellow for harvesting underway, and medium green for harvesting completed.

Disclaimer

No approval or endorsement of any commercial product by the National Institute of Standards and Technology (NIST) is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

References

AgGateway (2018). AgGateway Grain and Feed Council. Retrieved from http://www.aggateway.org/eConnectivityActivities/Councils/GrainFeed.aspx on June 8th, 2018.

AgGateway (2017) 2017 Annual Report. Retrieved from https://s3.amazonaws.com/aggateway_public/AgGatewayWeb/News/CommunicationsKit/AgGatewayAnnualReport_DRAFT_FINAL_singlePage_102317C_lowres.pdf on June 8, 2018.

AgGateway (2017A). AgGateway Teams Run Successful Grain Traceability Field Trial. Retrieved from http://www.aggateway.org/

Newsroom/2017PressReleases/AgGatewayTeamsRunSuccessfulGrainTraceabilityFieldTrial.aspx on June 8, 2018.

AgXML Consortium (2009). CommodityMovement, AgXML Standards,. Retrieved from http://www.agxml.org/Standards.aspx on June 8, 2018.

AnyLogic Company. (2018) Manufacturing Simulation. Retrieved from https://www.anylogic.com/manufacturing/ on June 8, 2018.

Badia-Melis, R., Mishra, P., and Ruiz-García, L. (2015). Food traceability: New trends and recent advances. A review. *Food Control 57*, 393-401. doi:10.1016/j.foodcont.2015.05.005

Bhatt, T., Buckley, G., McEntire, J.C., Lothian, P., Sterling, B., and Hickey, C. (2013). Making traceability work across the entire food supply chain, *Journal of Food Science, 78*, B21-B27.

Church, J. (May 2015), How General Mills is advancing a sustainable supply chain, www.generalmills.com [online] Retrieved from https://blog.generalmills.com/2015/05/how-general-mills-is-advancing-a-sustainable-supply-chain/ on June 8, 2018.

Donnelly, K., and Thakur, M. (2010). Food traceability perspectives from the United States of America and the European Union. *Økonomisk fiskeriforskning Argang, 19(*20), 1-8.

Dow (2018), From Cow to Cup. Retrieved from https://www.dow.com/en-us/water-and-process-solutions/markets/food-and-beverage/dairy/from-cow-to-cup on June 8, 2018.

Economic Research Service (ERS), U.S. Department of Agriculture (2017). Cost Estimates of Foodborne Illnesses. U.S. Department of Agriculture. Retrieved from https://www.ers.usda.gov/data-products/cost-estimates-of-foodborne-illnesses on June 8, 2018.

European Commission (2017). Food Traceability. European Commission [online]. Retrieved from https://ec.europa.eu/food/sites/food/files/safety/docs/gfl_req_factsheet_traceability_2007_en.pdf on June 8, 2018.

Flynn, D. (2014). USDA: U.S. Foodborne Illnesses Cost More Than $15.6 Billion Annually, *Food Safety News*, Retrieved from http://www.foodsafetynews.com/2014/10/foodborne-illnesses-cost-usa-15-6-billion-annually/ on June 8, 2018.

Hussain, M. and Dawson, C. (2013). Economic Impact of Food Safety Outbreaks on Food Businesses. *Foods 2*(4), 585–589.

National Institute of Standards and Technology. (2017) Agri-Food Manufacturing System and Supply Chain Integration, Retrieved from https://www.nist.gov/programs-projects/agri-food-manufacturing-system-and-supply-chain-integration on June 8, 2018.

Opara, L.U. (2003). Traceability in agriculture and food supply chain: a review of basic concepts, technological implications, and future prospects. *Journal of Food, Agriculture and Environment, 1*, 101-106.

Open Applications Group. (2016). Quality Content Work Group Update, Retrieved from https://oagi.org/Portals/0/Downloads/Meetings/2017_0412_Gaithersburg/2017-04-12%20OAGI%20QualityContentWGPlenaryAnnual.pdf on May 4th, 2018.

Schultz, H. (2018). FDA to delay enforcement on four FSMA rules, Retrieved from https://www.nutraingredients-usa.com/Article/2018/01/08/FDA-to-delay-enforcement-on-four-FSMA-rules on June 8, 2018.

Silver, B. (2011). *BPMN method and style: with BPMN implementer's guide, 2nd Ed*. Aptos, Calif: Cody.

Thakur, M. and Hurburgh, C. (2009). Traceability activities in the United States and the TRACE project. Proc. of the Final Trace conference – how to trace the origin of food?, Brussels, Belgium, December 2-3, 2009.

Thakur, M., Mosher, G., Brown, B., Bennet, G., Shepherd, H., and Hurburgh, C. (2009). Food Safety Series--Traceability in the Bulk Grain Supply Chain. *Resource Magazine, 16*, 20-22.

Trisotech. (2018). Trisotech Digital Enterprise Suite. Retrieved from https://www.trisotech.com/digital-enterprise-suite on June 8, 2018.

United Nations Global Compact. (2014). A Guide to Traceability: A Practical Approach to Advance Sustainability in Global Supply Chains, Retrieved from https://www.unglobalcompact.org/library/791 on June 8, 2018.

US Department of Defense (1998). *Department of Defense Modeling and Simulation (M&S) Glossary*, DoD 5000.59-M.

US Food and Drug Administration (2018). FDA Food Safety Modernization Act (FSMA), Retrieved from https://www.fda.gov/Food/GuidanceRegulation/FSMA/ on June 8, 2018.

Riddick, Frank H.; Wallace, Evan K.; Nieman, Scott; Tevis, Joe; Ferreyra, R. ”Implementing Grain Traceability Standards: CART and Simulation.” Paper presented at 2018 American Society of Agricultural and Biological Engineers (ASABE) Annual International Meeting, Detroit, MI, US. July 29, 2018 - August 01, 2018.

Compact 1.7 K Cryocooler for Superconducting Nanowire Single-Photon Detectors

V. Kotsubo, J.N. Ullom, and S.W. Nam

National Institute of Standards and Technology
Boulder, CO 80305, USA

ABSTRACT

State-of-the-art superconductor-based cryogenic detector systems are being installed at numerous research facilities worldwide and are achieving world-record sensitivities in a variety of applications. Implementation has been greatly facilitated by closed-cycle refrigeration. However, in many cases, cooling capacities of the refrigerators exceed requirements, at times by orders of magnitude, resulting in excessively large and cumbersome systems. The availability of more compact and lower power consumption systems should greatly facilitate further user acceptance. Toward this end, we are developing a compact 1.7 K closed-cycle pulse tube/Joule-Thomson hybrid cryocooler for superconducting nanowire single-photon detectors. A laboratory prototype consisting of the pulse tube cooler and the Joule-Thomson coldhead has demonstrated over 1.4 mW of cooling at 1.7 K. The Joule-Thomson compressor is under development and remains the single largest risk item in terms of reliability. The system, designed for low manufacturing costs, is projected to consume on the order of 250 W total power, including power for cooling fans, thermometry, and detector electronics, and to be mountable within a standard equipment rack.

INTRODUCTION

Applications for superconductor-based sensors and electronics have been steadily increasing over the past several years in diverse areas such as astrophysics/cosmology,[1] X-ray spectroscopy,[2] gamma-ray spectroscopy,[2] quantum information,[3] and photon science.[3] These systems operate at temperatures ranging from above 4 K to lower than 50 mK. Although these systems are developed by scientists with low-temperature expertise, end users typically have minimal cryogenic experience, and therefore acceptance has been greatly facilitated by push-button, closed-cycle cryocoolers. The majority of these systems depend on precooling with commercial Gifford-McMahon or valved pulse tube coolers, which were originally developed for applications requiring relatively large cooling capacities such as shield cooling in magnetic resonance imagers or cryopumping. As a result, for many, if not most of these systems, the size and power consumption of the system are orders of magnitude larger than necessary. For example, many transition edge sensor (TES) systems are cooled using adiabatic demagnetization refrigerators (ADRs) with cooling capacities of hundreds of nanowatts at temperatures below 100 mK. At the 4 K reject temperature for the ADR, the heat rejected is on the order of a milliwatt, yet precoolers with capacities on the order of a watt with a power draw of several kilowatts are used.

Kotsubo, Vincent; Ullom, Joel; Nam, Sae Woo. "Compact 1.7 K Cryocooler for Superconducting Nanowire Single-Photon Detectors." Paper presented at 20th International Cryocooler Conference, Burlington, VT, US. June 18, 2018 - June 21, 2018.

While there are advantages for using excessively large coolers such as rapid cooldown times and flexibility in cryostat design, future widescale applications of these types of systems will be greatly enhanced by reducing their size and power. Such applications include secure quantum encrypted communication links[3] that utilize superconducting nanowire single-photon detectors (SNSPD), and superconducting transition-edge sensor microcalorimeters for electron microscope microanalysis.[2]

Based on these considerations, we initiated the development of a compact cooler for an SNSPD detector system that will result in a form factor and low power draw of a standard rack- mountable electronics instrument. The cooler is a pulse tube/Joule-Thomson (PT/JT) hybrid, with the JT stage achieving 1.4 mW of cooling at 1.7 K, and the three-stage pulse tube providing cooling at 80 K, 25 K, and 10 K. To date we have successfully developed the pulse tube cooler and JT coldhead, whereas the JT compressor and associated gas decontamination is still under development.

We have previously published results on an early version of this cooler.[4] In this paper we present further details, plus improvements, including a modification of the pulse tube 10 K stage regenerator that increased the cooling capacity at 10 K by 35 % and reduced the minimum temperature from 7.8 K to 6.4 K; and a modification of the JT stage, which dropped the minimum temperature from 2.2 K to 1.7 K.

CRYOCOOLER DESCRIPTION

Design Goals

As discussed in a previous publication[4] the target application is a four-channel tungsten-silicide (WSi) SNSPD.[5] Estimated cooling loads on each of the cooler stages are presented in Table 1 along with the design goals. The cooling load on the JT stage arises primarily from low thermal conductance bias/readout coaxial cables, since SNSPD's dissipate negligible power and photons are coupled in through optical fiber. Our targeted cooling capacities included design margin on each stage. The 80 K stage radiation load estimate has considerable uncertainty due to unknown ambient temperature since the cryocooler will be enclosed in an equipment rack-mounted box, and because of the uncertain surface emissivity of the radiation shield.

The WSi SNSPD detectors require cooling to 1.25 K for maximum quantum efficiency, but this temperature is difficult to achieve with a ^{4}He JT refrigerator that utilizes a low power, compact compressor. A ^{3}He JT would reach this temperature and will be used for future units, but to facilitate development, we used ^{4}He and targeted 2.0 K since the identical vapor pressure if ^{3}He is used would result in 1.25 K. Operation at temperatures up to 2.5 K is acceptable if maximum quantum efficiency is not required.

The 10 K for the precooling temperature for the JT stage is a lower temperature than the thermodynamic optimum for this hybrid, but we opted for as low as reasonable precooling temperatures to reduce the high-pressure requirement for the JT loop, which in turn reduces stress on the JT compressor. We chose this approach because the JT compressor is the largest risk to reliability. Final test results shown below show that the actual operating temperature of this stage was below 7.7 K

Table 1. Estimated heat loads and design goals for a four channel SNSPD detector cooler.

Stage	Radiation (mW)	JT Precooling (mW)	Conduction (mW)	Total (mW)	Design Target (mW)
80 K	1000-2350	400	150	1550-2900	3000
25 K	5-12	28	7	40-47	100
10 K	0.2	2	1	3.2	5
2.0K	0	N/A	0.28	0.28	0.5

.

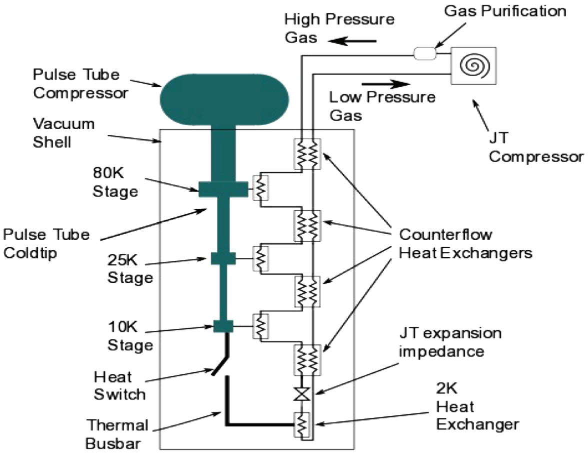


Figure 1. Operational schematic of the cryocooler.

Cooler Description

The cooler is shown schematically in Figure 1. The pulse tube is a three-stage, cooled inertance tube[6] configuration driven by a commercial resonant piston compressor at 35 Hz. Each stage is a U-tube configuration, resulting in an easy-to-integrate coldhead. Because the objective of the project is a producible, low cost cooler, the mechanical design sacrificed performance to reduce costs. The most impactful tradeoff was the use of standard sized stainless-steel tubing for the regenerator and pulse tube walls, rather than machined titanium, which is commonly used in aerospace coolers. Replacing the stainless-steel tubes with titanium tubes would result in 1.8 W, 0.07 W, and 0.001 W of additional cooling on the 80 K, 25 K, and 10 K stages, respectively.

The upper two-stage inertance tubes were dual diameters, whereas the third stage used only a single diameter because modeling indicated minimal improvement in performance with dual diameters. Regenerators used standard materials; the first stage used die-punched stainless-steel screens, with #150 mesh, 66 um wire diameter in the warm side and #400 mesh, 25.4 um wire diameter in the cold side. The second stage used #400 mesh, 25.4 um wire diameter stainless steel screens in the warm side and #400 mesh, 25.4 um wire diameter phosphor bronzes screens, flattened to produce 0.55 porosity, in the cold side. The third stage used high heat capacity microspheres. In the previous work reported in Ref. 4, 100 um diameter 50-50 erbium-praseodymium (ErPr) alloy spheres[7] were used. We have since replaced the lower half of the regenerator with 100 um diameter erbium-nickel (ErNi) spheres to improve cooling capacity below 10 K, and comparative tests results are presented below.

A critical mechanical design objective is geometries that minimize the effects of secondary flows within the pulse tube and regenerator, because these flows can seriously degrade performance. Most well-known are turbulence/convection in the pulse tube and streaming in regenerators.[8] Somewhat less appreciated are secondary flows in the displacer gap in Stirling coolers. A net circulation can flow through the displacer gap and return through the regenerator partially unregenerated and cause cooling loss unless there is sufficient lateral thermal conduction. Alternatively, because the displacer gap is dynamic, the local gap dimensions can vary during the operating cycle, causing local oscillating flows to not exactly reverse motion, again leading to unregenerated flow and cooling loss.

While it is well documented that, in large diameter regenerators, net circulation can lead to significant degradation in coolers,[8] we were concerned that even in small diameter regenerators, some level of circulation can exist, causing measurable cooling loss. In multi-stage coolers, circulation can be generated in the regenerators near the junctions between stages, where the phasing of the flows between the upper stage regenerator, lower stage regenerator, and buffer tube will necessarily lead to circulation in the regenerators if the upper and lower stage regenerators are in-line with each other because flows entering the regenerator can come from either the adjacent regenerator or the flow channel from the buffer tube. To mitigate this, we located the second stage regenerator such that the warm entrance was in the connecting channel between the first stage regenerator and first stage pulse tube rather than directly in-line with the first stage regenerator. Thus, the flow entering the cold end of the first stage regenerator always came from the channel. This arrangement was not possible for the junction between the second stage and third stage regenerators because the second stage regenerator packing access was through an attachment flange for the third stage regenerator, which was directly in line with the second stage regenerator. However, the small diameters of the second and third stages likely mitigated any effects of circulation. We have not validated whether this design feature improves cooling, but because the cooler performs as predicted by the model, it does not appear that there are any detrimental effects.

For the JT system, the very low cooling capacity requirement allowed a design driven by ease of fabrication rather than by thermodynamics. The stage nominally operates between a high pressure of 200 kPa and a low pressure of 3.2 kPa, with a flow rate of 0.6 mg/s. The thermodynamic power required to recompress the gas is only a few watts, which is such a small fraction of the total power budget that a highly thermodynamically efficient design is not required.

All four JT counterflow heat exchangers were simple tube-in-tube designs, with high-pressure gas flowing in the narrow annular space between the two tubes and the low-pressure gas flowing through the inner tube. This configuration simplifies fabrication, plus has a large internal surface area on the high-pressure return side to allow condensable contaminants to freeze out without plugging the flow passage. The expansion impedance was a 2 m long, 50 μm inner-diameter stainless steel capillary.

At the cold ends of the warmer three counterflow heat exchangers, a small heat exchanger consisting of fine mesh copper screen diffusion bonded to a copper body was used to heat sink the incoming gas to the pulse tube. The 1.7 K coldstage also had a small copper screen mesh heat exchanger for thermal contact to the load.

To mitigate particulate contamination, sintered stainless steel particulate filters were inserted in the high-pressure line prior to entering the cryostat, at the end of the heat exchanger at 80 K and just upstream of the JT expansion impedance. In addition, a small capsule of activated charcoal was placed in the high-pressure stream on the 80 K stage to adsorb any condensable contaminants.

To expedite cooldown of the JT coldstage from room temperature, a heat switch consisting of a small bar that clamped down on a small copper tab attached to the JT coldhead was used. The clamp was heat-sunk to the pulse tube 10 K stage and was actuated by pulling on a fine stainless-steel wire attached to the bar. The far end of the wire was fed through a bellows feedthrough through the room temperature vacuum flange to allow actuation of the heat switch.

In the original system,[4] both the inlet and outlet of the JT expansion capillary were housed in the same copper block and therefore at the same temperature. In this version, we separated the warm and cold ends of the capillary, and the result, presented below, was a lower base temperature of 1.7 K in comparison to the 2.2 K base temperature previously reported.

The JT compressor under development is a scroll compressor, selected because of the ability to achieve low suction pressures, compact size, and low per-unit cost.

MODELING RESULTS

Pulse Tube

The pulse tube was modeled using a widely used commercial Stirling and pulse tube software design package.[9] We relied heavily on the built-in numerical optimizing algorithm to design the cooler so, at least on paper, this is a highly efficient design. In this section we present modeling results, and although most of these model outputs are not easily measurable, we present them as an illustration of cooler behavior.

Table 2 presents top-level parameters of the pulse tube: frequency, charge pressure, oscillatory pressure amplitude P_1, compressor power, and cooling powers. The cooling powers identically match the design goals because those were set as constraints in the optimizer. The compressor power was the result of the optimization process that required its minimization.

Table 2. Modeled Top Level Pulse Tube Parameters

Frequency (Hz)	35
Charge Pressure (MPa)	1.5
Pressure Amplitude P_1 (MPa)	0.13
Compressor Power (W)	148
$\dot{Q}$ 1st stage @ 80 K (W)	3.00
$\dot{Q}$ 2nd stage @ 25 K (W)	0.10
$\dot{Q}$ 3rd stage @ 10 K (W)	5.0×10−3

Table 3. Modeled Compressor Loss

Total Compressor Power (W)	148
Delivered Acoustic Power (W)	117.5
Ohmic Loss (W)	17
Friction/Eddy Current Damping (W)	8.5
Compression Space Heat Transfer (W)	2.5
Seal Blowby (W)	2.5

The charge pressure of 1.5 MPa was driven by the limited regenerator heat capacity below 10 K, and this lower pressure sacrificed thermodynamic efficiency at the higher stages for increased thermodynamic efficiency at the lowest stage.

Table 3 lists the modeled losses for the pulse tube compressor, where the analysis used compressor parameters provided by the vendor. The ohmic loss is particularly small for this type of cooler because the compressor was significantly larger than necessary, with the capability of 600 W of input power.

Table 4 shows regenerator energetics. $\dot{E}_2$ is the second-order acoustic power, defined within the thermoacoustic framework[10] as:

$$\dot{E}_2 = \langle P_1 U_1 \rangle \tag{1}$$

where P_1 is the first order pressure amplitude and U_1 is the first order volume flow rate defined as:

$$U_1 = A_c v \tag{2}$$

where A_c is the cross-sectional flow area, v is the local spatial averaged velocity, and the brackets denote time averaging. The modeling includes higher harmonics, but the acoustic power associated with the fourth-order term is negligible so only the second-order acoustic power is presented. $\dot{H}_{\text{cold}}$, $\dot{Q}_{\text{screen}}$ and $\dot{Q}_{\text{tube}}$ are the enthalpy, screen conduction, and tube wall conduction losses. The enthalpy and screen conduction loss were taken at the cold end of the regenerator because the model couples these two losses internally within the regenerator. As a figure-of-merit, we define the acoustic power loss ratio as:

$$\varepsilon_{\dot{W}_2} = \left(\frac{\dot{E}_{2,cold}}{\dot{E}_{2,hot}}\right) \Bigg/ \left(\frac{\dot{E}_{2,cold,ideal}}{\dot{E}_{2,hot,ideal}}\right) = \left(\frac{\dot{E}_{2,cold}}{\dot{E}_{2,hot}}\right) \times \left(\frac{T_{hot}}{T_{cold}}\right) \tag{3}$$

where, for an ideal lossless regenerator using an ideal gas, the ratio of the acoustic powers at the ends of the regenerator scale as the ratio of the temperatures. The subscripts *hot* and *cold* refer the hot and cold ends of the regenerator. The acoustic power loss ratio therefore indicates the remaining fraction of the ideal acoustic power not consumed by irreversibilities such as viscous loss, within the regenerator.

The second figure-of-merit is the ratio:

$$\varepsilon_{\dot{H}} = \frac{\dot{E}_{2,cold} - \dot{H}_{cold}}{\dot{E}_{2,cold}} \tag{4}$$

which represents the fraction of the gross cooling power remaining after subtracting the enthalpy flow loss. The results of Table 4 indicate that acoustic power loss and enthalpy loss dominate, with

Table 4. Modeled Regenerator Energetics.

	$\dot{E}_{2,\text{hot}}$ (W)	$\dot{E}_{2,\text{cold}}$ (W)	$\dot{H}_{\text{cold}}$ (W)	$\dot{Q}_{\text{screen}}$ (W)	$\dot{Q}_{\text{tube}}$ (W)	ε_{W_1}	$\varepsilon_{\dot{H}}$
1st stage	111.6	19.0	5.80	0.53	1.32	0.66	0.69
2nd stage	3.46	0.97	0.50	0.026	0.048	0.90	0.48
3rd stage	0.086	0.031	0.01	8.0×10−5	0.0013	0.90	0.68

sizable acoustic power loss occurring in the first stage regenerator, while enthalpy losses occur in all three regenerators but is the largest in the second stage. Part of the reason for high second stage enthalpy loss is that it is operating near its minimum temperature of about 21 K. The high acoustic power efficiencies of the second and third stages in comparison to the first stage suggests that, by using commercially available screen meshes, the hydraulic diameters are larger than optimum, resulting in higher enthalpy losses but lower acoustic losses.

Table 5 shows the energetics of the thermal buffer tubes. The efficiency is:

$$\varepsilon_{TBT} = \frac{\dot{H}_{cold}}{\dot{E}_{2,cold}} \tag{5}$$

In the model, the enthalpy loss and tube wall thermal conduction loss are coupled so that the losses shift between the two along the length of the tube, so the enthalpy and tube conduction are taken at the cold end of the tube. The efficiencies are reasonably high, with one of the contributing factors being the low fluid displacement amplitudes, in the range of 10 % of the total length.

Joule-Thomson Modeling

Because of the very low flow rates and low thermodynamic compression power, the JT loop did not require a high degree of optimization, so counterflow heat exchangers were designed for ease of fabrication. They were modeled using a simple NTU analysis[11] using temperature-averaged fluid properties, along with a separate calculation of thermal conduction down the heat exchanger tubes. There was considerable leeway in the design, allowing heat exchanger lengths to vary by a factor of two without significantly affecting performance. The main consideration was minimizing the demand on the JT compressor, so the system was designed for the lowest reasonable high-side pressure to minimize stresses on bearings and backleakage through seals. Final flow and pressure requirements for the compressor were determined from test data.

TEST RESULTS

Pulse Tube Results

Prior to integration with the JT stage, the pulse tube was tested as a stand-alone unit. It was installed in a laboratory test dewar, with the warm-end temperature controlled to 310 K using temperature-regulated circulating water to simulate the anticipated higher temperature environment when the unit is installed within a fan-cooled equipment rack. The test dewar was not temperature controlled; its temperature varied considerably because the laboratory temperature was unregulated. As a result, there was significant run-to-run variability in the radiation load on the 80 K stage. The PT compressor and first-stage inertance tube were also not temperature regulated, which also led to some first-stage performance variability.

Modeling of the inertance tubes is generally not expected to be accurate because of turbulence, sensitivity to tube diameter, and uncertainty in the heat sink temperature, so an initial tuning of the inertance tube was required. Because inertance tubes for small capacity coolers are effectively low-*Q* quarter-wavelength acoustic resonators, the tubes were tuned by sweeping the drive frequency of the compressor and searching for the frequency of the maximum cooling capacity of each stage. The lengths of the inertance tubes where trimmed or extended by the ratio of this frequency to the resonant frequency of the compressor. The tubes were designed with as low a Q as possible to reduce sensitivity to length and temperature.

Upon completion of this tuning, the first characterization test was to simply measure the

Table 5. Energetics of the thermal buffer tube.

	$\dot{E}_{2,\mathrm{cold}}$ (W)	$\dot{H}_{\mathrm{cold}}$ (W)	ε_{TBT}	$\dot{Q}_{\mathrm{tube}}$ (W)
1st stage	14.2	10.4	0.73	0.42
2nd stage	0.82	0.68	0.83	0.043
3rd stage	0.021	0.017	0.81	7.4×10−4

cooling powers for each stage at their nominal design temperatures of 80 K, 25 K, and 10 K with 150 W compressor power. The results are tabulated in Table ww, with a drive frequency of 35 Hz and charge pressure of 1.5 MPa. The 10 K stage regenerator for this test consisted of all ErPr spheres. The measured loads are in addition to an unknown amount of radiation loading so the actual cooling capacity is somewhat higher on the 80 K and 25 K stages than shown in the table. Reasonable estimates of the radiation load suggest that the cooling power of the 80 K stage is in excess of the design point goal of 3.0 W.

Figures 2a and 2b show the cooling capacities of the 80 K stage and 25 K stage as a function of temperature. Each load curve was taken while holding the other two stages at constant temperature. These plots illustrate that the 80 K stage handles loads on the order of watts, while the 25 K stage handles loads on the order of tenths of watts.

Figure 3 show load curves for the 10 K stage and compares the performance of the all ErPr regenerator and the regenerator with ErPr in the warm half and ErNi in the cold half, with data for ErPr/ErNi regenerator combination shown at three different charge pressures. In all cases the ErPr/ErNi regenerator outperforms the all-ErPr regenerator, except at temperature above 10 K with 1.25 MPa charge pressure for the ErPr/ErNi regenerator. The ErPr/ErNi combination also consistently reached a minimum temperature in the 6.3 K to 6.5 K range while the minimum temperature of the all ErPr regenerator reached a minimum temperature slightly below 8 K. These lower temperatures were expected because of the lower heat capacity of ErPr compared to ErNi at lower temperatures. These data were taken with no applied load to the upper two stages such that the upper stages ran colder than the data in Table 6, which resulted in higher 10 K cooling capacity.

Joule Thomson Results

Because the JT compressor is still under development, final tests on the integrated JT-pulse tube coldhead was conducted open loop using a regulated compressed gas storage bottle as a supply and a vacuum pump vented to atmosphere on the return. The low side pressure was controlled by adjusting a valve in front of the vacuum pump.

Figure 4 shows load curves at high side pressures of 0.2 MPa, 0.15 MPa, 0.125 MPa, and 0.1 MPa. A maximum of 1.4 mW was produced at the minimum temperature of about 1.7 K. The cooling capacity increase at lower temperatures is from increased mass flow through the capillary from higher gas density. Operating at 0.1 MPa does not provide sufficient cooling to cool down the JT stage from the initial temperature of around 7 K.

Table 6. Simultaneously measured cooling capacities at all three stages with 150 W compressor power.

Stage	Temperature (K)	Capacity (W)
1	80	2.77
2	25	0.080
3	10	5.0×10−3

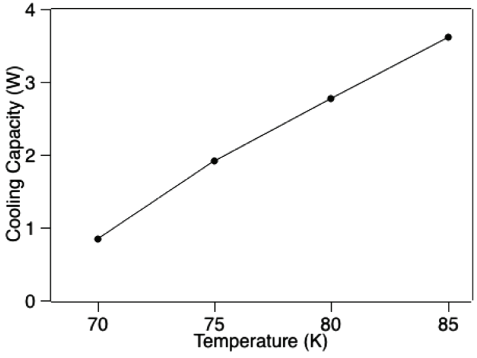


Figure 2a. Load curve for the 80 K stage

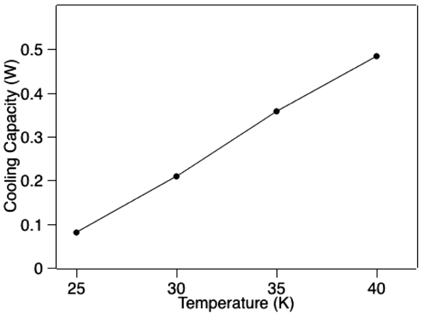


Figure 2b. Load curve for the 25 K stage

In the previous publication we reported instability at the lowest temperatures.[4] In the present setup, the instability occurs only when the temperature of the upstream side of the capillary drops below the liquefaction temperature corresponding to high side pressure. We attribute this to liquid in the low-pressure pumping line reaching the cold end of the JT counterflow heat exchanger, which then liquefies the fluid entering the capillary. Although we do not specifically know why this causes the instability, if we temperature-regulate the coldstage slightly above the minimum temperature, no liquid accumulates, which stabilizes the cold stage temperature.

The return pressure in these tests was as low as 650 Pa. The return pressure does not affect the load curves except that lower temperatures are reached with lower return pressures as expected from the saturation curve.

During these tests, the load on the pulse tube from circulating ^{4}He would cause the third stage temperature to rise slightly. However, in all situations the temperature remained below 7.7 K, indicating that Nb-Ti wire can be used for electrical leads between the 10 K stage and the JT stage.

Cooldown Time

The system cooled to base temperature within 24 h. The 80 K pulse tube stage reached close to its minimum temperature in about 4 h, but both the 25 K and 10 K stages took significantly longer because of their low cooling capacities and large thermal mass loads associated with the compliance tanks and JT components. Once the PT reached its bottom temperature, the JT stage would cool rapidly to its base temperature, within a few minutes if 0.2 MPa high-side pressure was used.

TOTAL POWER PROJECTION

Test data described above show that the design point cooling powers for the pulse tube require a nominal 150 W of compressor power. We have tested a variety of commercial off-the-shelf power electronic modules for driving the compressor and they generally had conversion efficiencies of better than 90 % power so the pulse tube will require a total power draw of about 165 W. Although the JT compressor is still under development, preliminary tests indicate that the power draw by the compressor and drive electronics will be about 50 W. If the power for cooling fans and thermometry/diagnostic instrumentation is included, the final total power draw may be well less than 250 W. We also evaluated power consumption if we used a smaller, commercially available, pulse tube compressor to reduce system weight, with the result that another 30 W of compressor power will be required.

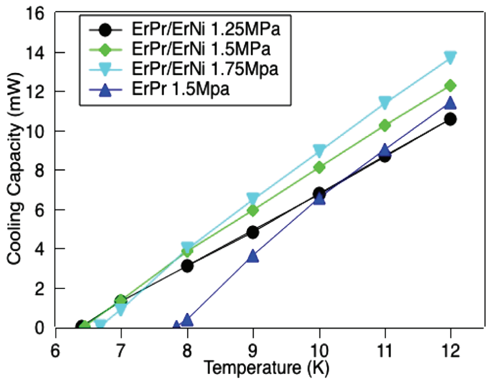


Figure 3. Load curves for the 10 K stage comparing the all ErPr regenerator to the ErPr/ErNi regenerator at pressures of 1.25 MPa, 1.5 MPa, and 1.75 MPa.

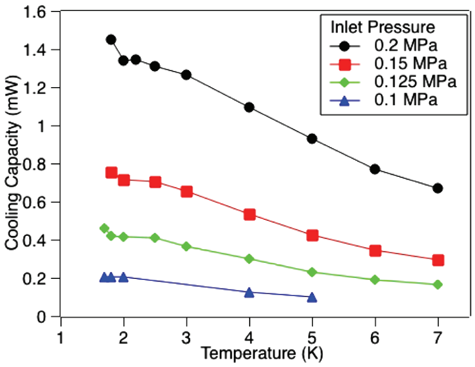


Figure 4. Load curves for the JT stage at high side pressures of 0.1 MPa, 0.125 MPa, 0.15 MPa, and 2 MPa.

EXPANDABILITY

These test results are for one operating condition for the pulse tube and one flow impedance for the JT stage. For low heat load applications, the flow impedance can be increased, reducing the mass flow, which would both reduce the power required for both the JT compressor and pulse tube compressor. Conversely, there is substantial upside in the cooling power capability of the JT stage. The JT impedance can be reduced, which would result in higher mass flow rates and higher cooling powers. The pulse tube has sufficient capacity to accommodate higher JT mass flow rates, so capacities approaching 5mW are possible, although this will require development of a larger capacity JT compressor.

Lower temperatures using ^{4}He are possible but will require a combination of reworking the JT counterflow heat exchanges to reduce pressure drop, a JT compressor with lower suction pressure, and superfluid film creep mitigation. Direct substitution with ^{3}He should result in temperatures below 1 K.

CONCLUSION

We have demonstrated a compact, low power pulse tube/JT hybrid cooler that reaches a base temperature of 1.7 K with up to 1.4 mW of cooling. We project that an SNSPD system using this cooler will have a total power draw of about 250 W and the entire unit will be rack mountable in a standard equipment rack. The component with the highest risk to reliable operation, the JT compressor, is presently under development.

ACKNOWLEDGMENT

This work was supported by the Laboratory of Telecommunications Sciences agreement No. EAO222214 and by the National Institute of Standards and Technology's Physical Measurement Laboratory to advance the goal of "Invisible Cryogenics."

Contribution of a U.S government agency, not subject to copyright.

REFERENCES

1. J. Hubmayr, J.E. Austermann, J.A. Beall, D.T. Becker, B. Dober, S.M. Duff, J. Gao, G.C. Hilton, C.M. McKenney, J.N. Ullom, J. Van Lanen, M.R. Vissers, "Low Temperature Detectors for CMB Imaging Arrays," *J. Low Temp. Phys*, to be published.
2. Ullom J.N. and Bennett, D. A., "Review of Superconducting Transition Edge Sensors for X-Ray and Gamma-Ray Spectroscopy," *Superconductor Science and Technology*, vol. 28, 084003 (2015).
3. Hadfield, R.H., "Single-Photon Detectors for Optical Quantum Information Applications," *Nature Photonics,* vol. 3 (2009), pp 696-705.
4. Kotsubo, V., et. al. "Compact 2.2K Cooling System for Superconducting Nanowire Single Photon Detectors," *IEEE Trans. Appl. Superconductivity,* vol. 27, Issue 4 (2017), pp 1-5.
5. F. Marsili, V.B. Verma, J.A. Stern, S. Harrington, A.E. Lita, T. Gerrits, I. Vayshenkar, B. Baek, M.D. Shaw, R.P. Mirin, and S.W. Nam, "Detecting Single Infrared Photons with 93% System Efficiency," *Nature Photonics* 7 (2013), pp 210-214.
6. B. Wang, L.Y. Wang, J.K. Zhu, J. Chen, Z.P. Li, Z.H. Gan, L.M Qiu, "Study of Phase Shifting Mechanism of Inertance tube at Low Temperatures," *Cryocoolers 17,* ICC Press, Boulder, CO (2012), pp 169-177.
7. A. Kashani, B.P.M Helvensteijn, J.R. Maddocks, P. Kittel, J.R. Feller, K.A. Gschneidner, V.K. Pecharsky, A.O. Pecharsky, "New Regenerator Materials for use in Pulse Tube Coolers,"*Cryocoolers 12*, Kluwer Academic/Plenum Publishers, New York (2003), pp 475-480.
8. J. H. So, G.W. Swift, and S. Backhaus, "An Internal Streaming Instability in Regenerators," J. Acous. Soc. Am. 120 (2006), pp 1898-1909.

Kotsubo, Vincent; Ullom, Joel; Nam, Sae Woo. "Compact 1.7 K Cryocooler for Superconducting Nanowire Single-Photon Detectors." Paper presented at 20th International Cryocooler Conference, Burlington, VT, US. June 18, 2018 - June 21, 2018.

9. Sage, Gedeon Associates, Athens, OH. Certain commercial equipment and software, are identified to describe the subject adequately. Such identification does not imply recommendation or endorsement by the NIST, nor does it imply that the equipment identified is necessarily the best available for the purpose.

10. G. W. Swift, *Thermoacoustics*, Springer International Publishing, (2017).

11. W. M Kays and A. L. London, *Compact Heat Exchangers*, McGraw-Hill (1984), pp 19-29.

Structure Vulnerability to Firebrands from Fences and Mulch

Kathryn M. Butler, **Erik L. Johnsson**, and **Wei Tang**,
National Institute of Standards and Technology, Gaithersburg, Maryland

Abstract—Fences and mulch contribute to the spread of wildland-urban interface fires, acting both as ignition targets and as sources that may ignite nearby objects through direct flame contact and firebrand generation.

This paper presents the findings from outdoor experiments that investigated the spread of fire through firebrand spotting from fences and mulch beds near a structure in a wind field. A fence section, mulch bed, or combination was arranged perpendicular to a small structure, with a large fan directing wind toward the structure. After ignition, data were collected on firebrand spotting time, time for the spot fire to reach the wall, and flame spread rate over the fence and mulch bed.

Fence type, wind speed, and type of mulch were found to affect firebrand spotting. For fence and mulch combinations, spotting usually occurred within 7 minutes. Time to spotting generally decreased with increasing wind speed. Two parallel fence panels burned significantly more intensely, and spotted more quickly, than a single fence panel. In the absence of mulch, spotting from fences often occurred more slowly or not at all. The combination of a fence and a mulch bed decreased the time to spotting over either the mulch bed or the fence alone.

Keywords: fence, firebrand, mulch, structure vulnerability, wildland-urban interface (WUI) fire

INTRODUCTION

Wildland-urban interface (WUI) fires threaten an estimated 70,000 communities, 46 million homes, and 120 million people within the United States (ICC and NARCD Councils 2013). In recent years, the United States has been losing on the order of 3,000 homes per year, and the costs are rising, with $14 billion spent in 2009 alone on fire suppression and damages.

Firebrands generated by a wildfire are carried by the wind and may ignite fires in a community far downstream of the fire front. After the fire has reached the community, firebrands generated from burning combustible objects near a structure contribute to the firebrand assault. Postfire investigations have demonstrated that firebrands are a major contributor to structure losses in WUI fires.

Fences and mulch are common contributors to the spread of WUI fires within WUI communities. They act both as ignition targets and as sources that may themselves ignite nearby objects through direct flame contact and firebrand generation. The linear nature of fences gives them the capability of spreading fire over long distances. In a study of the 2011 Tanglewood Complex Fire near Amarillo, Texas performed by the National Institute of Standards and Technology (NIST) (Maranghides and McNamara 2016), 2.4 km of fences within a community of 25 homes were found to be damaged or destroyed. Combustible fences also contributed to fire spread in the 2012 Waldo Canyon Fire in Colorado (Maranghides et al. 2015). Firefighters were documented removing fences as part of their defensive strategy to contain this fire, reducing resources allocated to suppression.

In: Hood, Sharon; Drury, Stacy; Steelman, Toddi; Steffens, Ron, tech. eds. The fire continuum—preparing for the future of wildland fire: Proceedings of the Fire Continuum Conference. 21-24 May 2018, Missoula, MT. Proc. RMRS-P-78. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 358 p.

Papers published in these proceedings were submitted by authors in electronic media. Editing was done for readability and to ensure consistent format and style. Authors are responsible for content and accuracy of their individual papers and the quality of illustrative materials. Opinions expressed may not necessarily reflect the position of the U.S. Department of Agriculture.

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018.

The goal of this work is to improve our understanding of the mechanisms by which fences and other combustible landscaping elements can transport fire to a home, including exposure to wind-driven firebrands. The results will be used to improve codes and standards, which in turn will provide guidance to homeowners and firefighters.

EXPERIMENTAL DESIGN

Setup

To investigate the spread of fire through firebrand spotting, a series of field experiments are being performed on fences, mulch beds, woodpiles, and other combustible landscaping elements arranged in front of a structure in a wind field.

Figure 1 shows a schematic of the experimental setup for fences and mulch beds. A wind machine, consisting of an airboat fan mounted on a trailer, was aimed toward a small structure. A flow straightener directed the wind downward slightly so the wind field would reach the ground and the base of any combustible object being tested. A fence section, with or without a mulch bed beneath, was arranged perpendicular to the wall of the structure. The fence section was 2.44 m long and 1.83 m high for privacy fences or 1.22 m high for lattice fences, attached to 0.09 m × 0.09 m pine (*Pinus* spp.) posts at each end. The fence or mulch bed was placed in contact with the wall of the structure or separated from it by some fixed distance.

To study the ability of firebrands to threaten a house, a target pan of hardwood mulch that was 0.46 m wide was arranged at the base of the structure wall. This mulch bed served as a surrogate for any combustible material next to a house. Because of its rough texture, any firebrands landing on this surface tended to stay in place.

Before being arranged in steel pans, the mulch was dried to a moisture content between 6 and 7 percent, as measured by a moisture analyzer. The mulch beds were prepared by filling the pans with an even layer of mulch and compressing it lightly by foot. The mulch beds were 0.05 m thick except for cases in which the mulch beds were reduced to half thickness (0.025 m).

The wind field was monitored by a set of bidirectional probes just upwind of the end of the fence. The ambient wind speed and direction were measured by

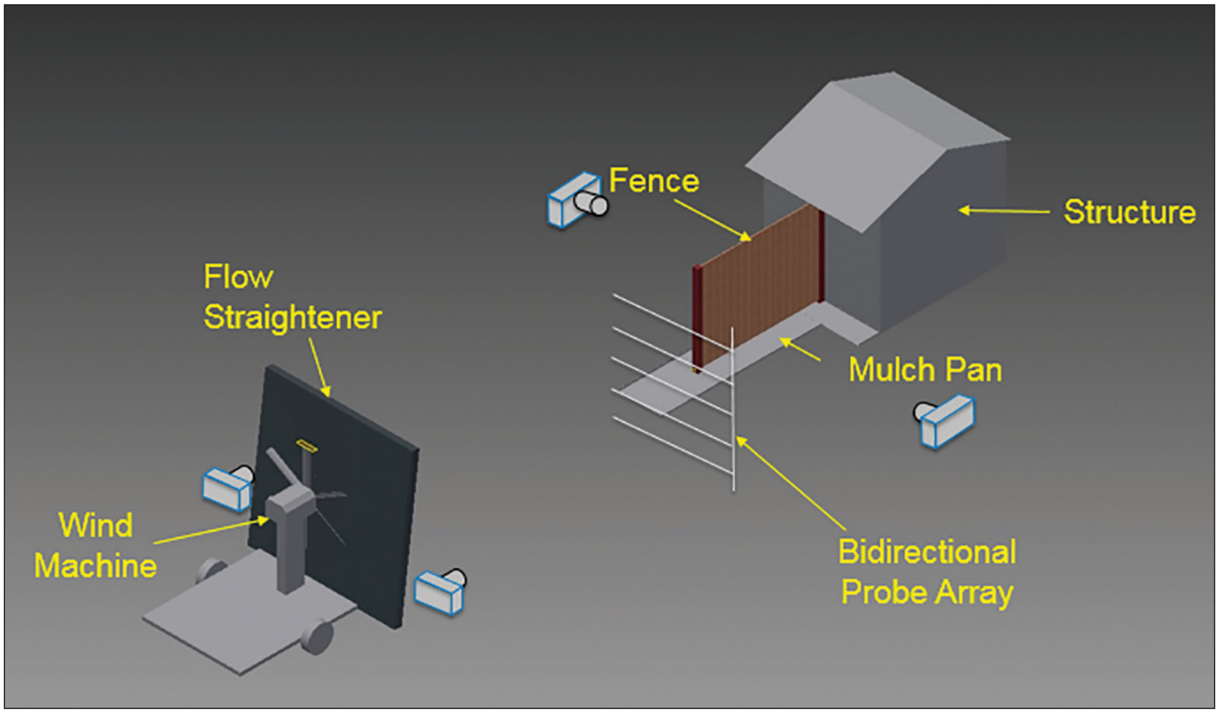


Figure 1—Experimental setup.

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018.

an anemometer mounted on a shed away from the experimental setup, and the ambient temperature was measured with a thermocouple near the test setup. Four video cameras monitored the experiment from right and left sides and from each side of the fan.

The ambient wind speed was required to be less than one-third the nominal wind speed in order to carry out the experiment. Under these conditions, the impact of the ambient winds on the wind field generated by the fan was minimal.

Procedure

A propane burner was used to ignite the fence or mulch bed at the base of the fence farthest from the structure. After 90 seconds, when the fire was judged to be self-sustaining, the fan was turned on, a timing clock was started, and the winds were brought to the speed required by the experiment. The experiment ended when a fire in the mulch bed at the base of the structure reached the wall and after fire had also reached the end of the fence or fence mulch bed. Flames at the wall from spot fires were extinguished if the fire had not yet spread over the entire length of the fence with mulch bed. At the end of the test, the clock was reset and all fires were extinguished with a water hose.

Uncertainties

The measurements of wind speed, distances, and times discussed in this paper each have uncertainties associated with them. Uncertainties generally consist of several components, which are grouped into two categories according to the method used to estimate their value. Type A uncertainties are evaluated by statistical methods, and type B uncertainties are evaluated by other means, often based on scientific judgment using all available relevant information (Taylor and Kuyatt 1994). Type B uncertainties are evaluated by estimating lower and upper limits a_- and a_+, such that the probability that the value lies in the interval a_- to a_+ is essentially 100 percent. If the value is equally probable to lie anywhere within the interval, the best estimate is $(a_+ + a_-)/2$, with standard deviation $u_j = a/\sqrt{3}$, where $a = (a_+ - a_-)/2$. Once all components have been estimated by either type A or type B analysis, they are combined using the square root of the sum of the squares (RSS) method to yield the combined standard uncertainty (estimated standard deviation), u_c. Finally, expanded uncertainties are given by $\pm\, ku_c$, where $k = 2$ is the coverage factor for a confidence level of 95 percent.

Table 1 shows the components of uncertainty for the measurements given in this paper. The extensive data collection on wind speed enables the evaluation of type A uncertainties; most other uncertainties on this list are type B, either estimated through scientific judgment or obtained from the literature.

Wind speed uncertainties involve the bidirectional probe design and the measurement statistics from the wind field. A paper by McCaffrey and Heskestad (1976) states that velocities are estimated within ±10 percent provided the approach flow direction is within approximately 50° of the probe axis. Since the variability of the fan was greatest at the lowest setting, which was close to the idle speed, statistical analysis was carried out for each wind speed level separately. Repeatability was calculated as the standard deviation of the average wind speed for each experiment (the average from five probes in the central wind field) from the average wind speed overall. The random component reflects the fluctuations in measured wind speed due to turbulence, and was calculated by the root-mean-square of the standard deviations of wind speed over all experiments at each wind speed level.

The time to spotting required identification of two events in videos taken by the camera positioned to the right or left side of the experiment. The first event was the time at which the fan was turned on. The engagement of the fan engine could be determined very accurately, although it should be noted that the wind speed was adjusted for up to 20 seconds afterwards before reaching a steady state value. The second event was the ignition within the target mulch bed of the first spot fire that eventually reached the wall of the structure. After identification in one of the videos, this spot fire was tracked backwards in time to the point at which the first sign of smoke could be detected in the mulch, which could be defined within an estimated ± 5 seconds. These sources of uncertainty are likely dwarfed, however, by the repeatability of time to spotting for multiple tests under the same conditions and the random nature of firebrand generation and ignition processes, neither of which is available for this study.

Table 1—Uncertainty in experimental data.

Measurement	Component standard uncertainty, ± u_j	Combined standard uncertainty, ± u_c	Total expanded uncertainty, ± $2u_c$
Wind speed			
Calibration	±10%		
Repeatability*			
6 m sec-1	±11%		
10 m sec-1	±6%		
14 m sec-1	±8%	6 m sec-1: ±24%	6 m sec-1: ±48%
Random*		10 m sec-1: ±15%	10 m sec-1: ±31%
6 m sec-1	±19%	14 m sec-1: ±15%	14 m sec-1: ±30%
10 m sec-1	±10%		
14 m sec-1	±8%		
Time to spotting			
Fan on	±1 sec		
Smoke detected	±3 sec	> ±3 sec	> ±6 sec
Repeatability*	Unknown		
Random*	Unknown		
Separation distance			
Placement	±0.002 m	±0.004 m	±0.008 m
Adjustment	±0.003 m		
Mulch thickness			
Variability	±0.005 m	±0.005 m	±0.010 m
Target mulch bed width			
Variability	±0.01 m	±0.01 m	±0.02 m

* Type A uncertainty (evaluated by statistical means). All other uncertainties are type B (evaluated by other than statistical means).

The separation distance between the fence or mulch bed and the wall of the structure was established by using a tape measure to adjust the location of the fence with mulch bed to the desired position. Sources of uncertainty include the placement of the tape measure and the ability to adjust the position of the fence with mulch bed accurately.

The mulch bed thickness varied over its surface due to the nature of the mulch as overlapping particles whose individual thicknesses are an appreciable fraction of the thickness of the mulch layer. The mulch bed thickness depended on the evenness of the spreading over the mulch bed and the uniformity of the compaction.

The target mulch bed at the base of the shed was not confined by a lip on the outside edge facing the fence, allowing firebrands to land on the target mulch without needing to clear a height. The width of the target mulch bed thus varied over its length.

Variations of dimensional values were estimated using scientific judgment.

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018.

Experimental Combinations

During 2016 and 2017, 111 experiments were carried out in the configuration just described. Figure 2 shows the distribution of experiments that have been performed on a variety of combinations of fences and mulch, at four separation distances from 0 m to 1.8 m, and at three wind speeds of 6 m second^{-1}, 10 m second^{-1}, and 14 m second^{-1}. The fences include privacy fences constructed of western redcedar (*Thuja plicata*) and vinyl and lattice fences constructed of redwood (*Sequoia* spp.) and pine. The mulches include shredded hardwood mulch at two thicknesses, pine bark mulch, and pine straw mulch.

RESULTS

Burning Characteristics for a Privacy Fence with Mulch Bed

Figure 3 shows an image from a typical experiment with a western redcedar fence sitting in a bed of shredded hardwood mulch. In this experiment, the wind speed was 10 m second^{-1} and the separation distance between the end of the fence and the small structure was 1.8 m. This image shows the conditions at about 4.5 minutes after the fan was turned on following ignition of the fence and mulch. At this point firebrands had ignited spot fires in the mulch bed at the base of the structure at several locations. A few spot fires can be seen at the front edge of the mulch bed, in addition to one close to the wall. The fence itself was burning along its entire length, although discoloration and other signs of deterioration show that the fire has remained low on the fence, not even reaching half of its height.

Other phenomena apparent from the video itself include pieces of mulch that moved out of the bed under the structure and rolled on the pavement toward the fence or the sides of the experiment. The smoke and flames from the combination of fence and mulch bed generally extended toward the structure, while smoke and flames from the mulch bed at the base of the structure extended toward the fence. These observations are consistent with a horseshoe vortex that occurs at the base of a structure at right angles to a flow stream (Martinuzzi and Tropea 1993). Figure 4 shows this feature in a model of the experimental

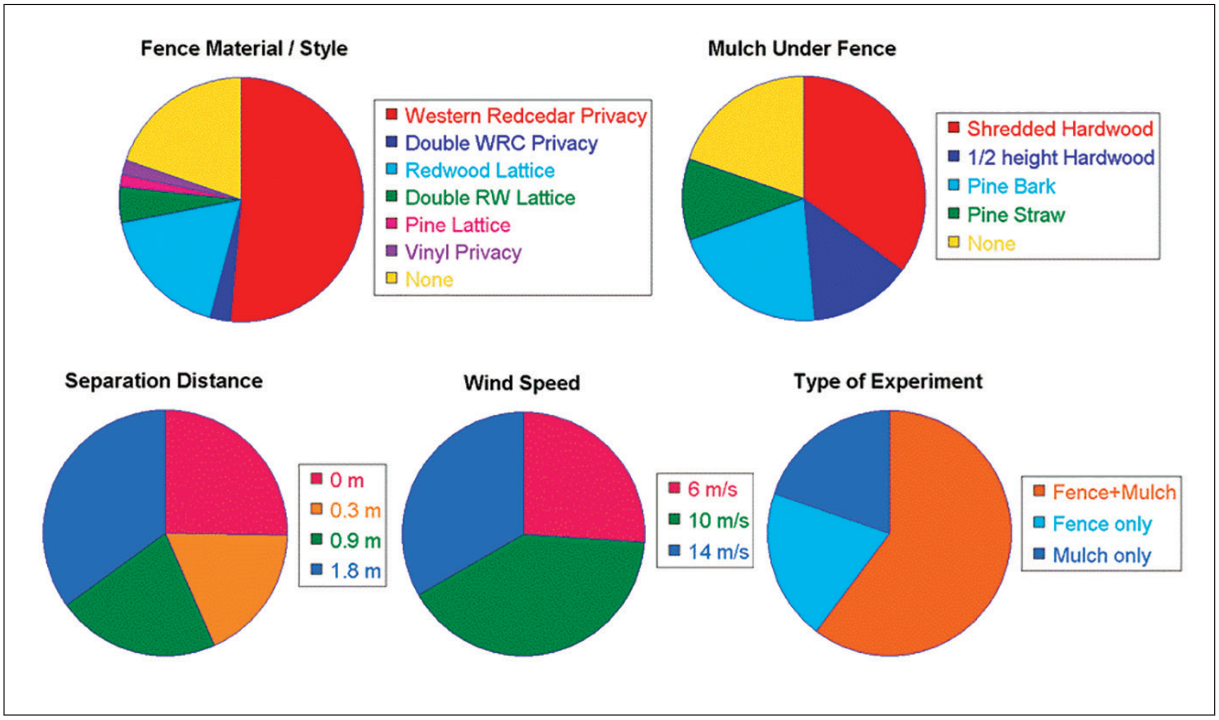


Figure 2—Distribution of experiments by fence type, mulch type, separation distance from wall, wind speed, and experiment type (fence with mulch beneath, fence only, or mulch only). WRC = western redcedar.

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018.

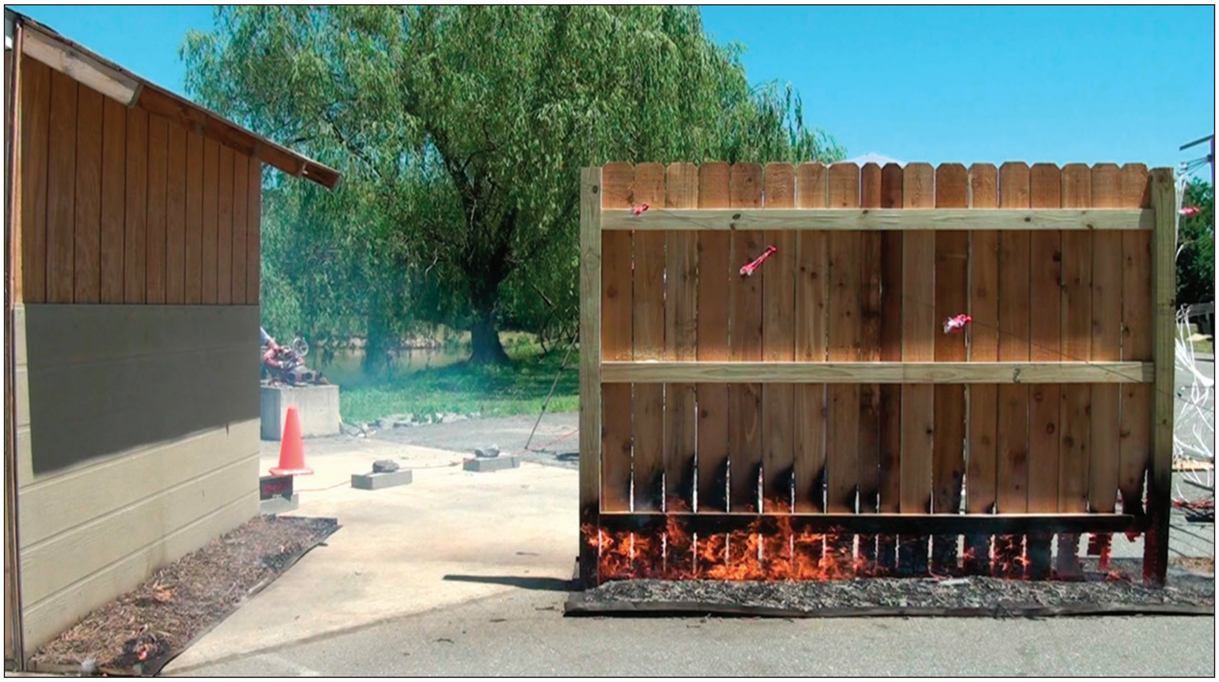


Figure 3—Image from video of western redcedar fence combined with shredded hardwood mulch.

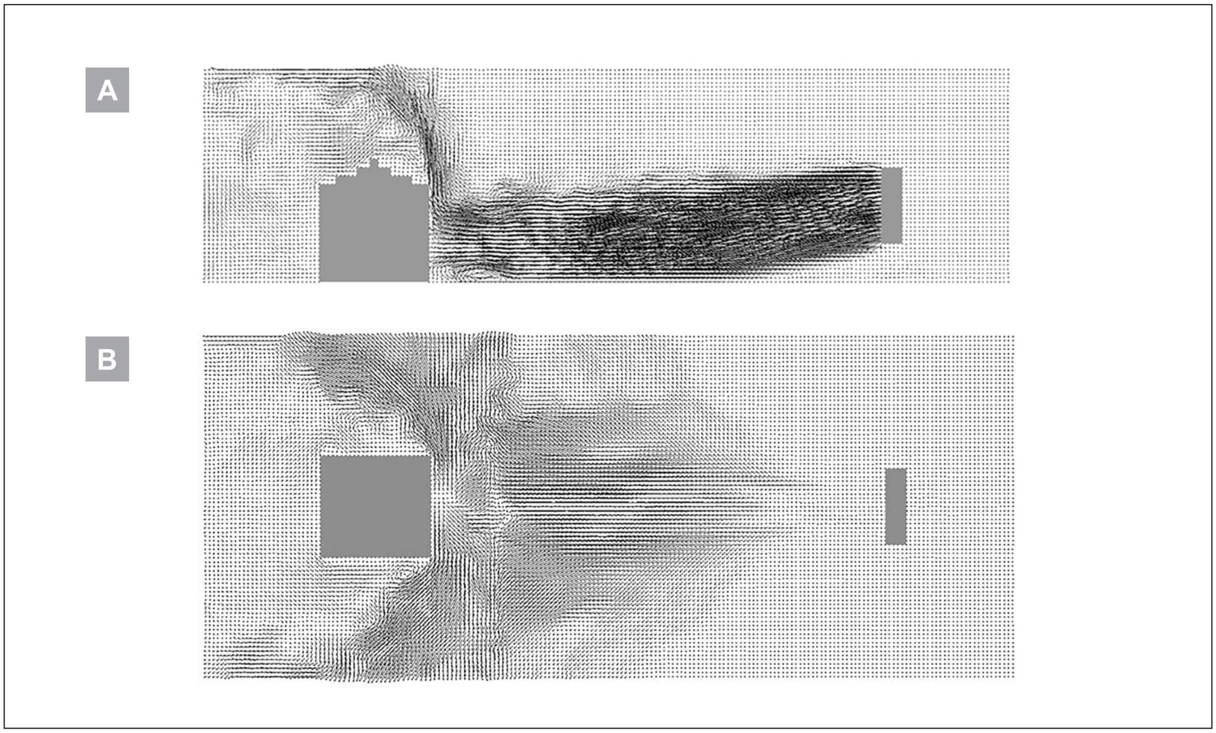


Figure 4—Instantaneous flow field from Fire Dynamic Simulator model, showing (A) side view and (B) top view in a plane close to the ground. Note the recirculation zone near the base of the structure wall.

setup using the NIST Fire Dynamic Simulator (FDS) (McGrattan et al. 2013). The vortex causes particles and smoke in the mulch bed at the base of the structure to be generally transported away from the wall and to the sides. Firebrands that are lofted, however, may be carried along the top of the vortex or drop out of the flow directed over the top of the structure, to be deposited at the base of the structure where they can ignite combustible materials close to the wall.

Spotting Time

This set of experiments represents a survey of the effects of fences and mulch on the spread of fire to a structure, directly or through firebrands and in a variety of conditions. Few experiments have been replicated, and many phenomena involved in firebrand spotting, such as generation of firebrands and ignition processes, are stochastic in nature. The analysis of this data was therefore based on uncovering trends and on discovering different modes of behavior, rather than on quantitative results.

One of the simple measures that has been determined from the video records is the length of time between turning on the fan and ignition within the target mulch bed at the base of the structure of the first spot fire that eventually reached the wall. Ignition was detected by the first sign of smoke.

Effects of Wind Speed and Separation Distance

In figure 5, the time to spot is plotted as a function of the nominal wind speed of the fan. The experiments represented in this plot are the 22 experiments performed on mulch alone and 67 experiments on fence and mulch combinations that are included in the pie charts of figure 2.

Figure 5 demonstrates several trends. First, the time to spot generally decreases as a function of wind speed. Second, the spotting times for fences in combination with mulch beds tend to be shorter than for mulch alone. Third, the spotting times are on the order of minutes. For 6 m second-1 winds, spotting occurs in 30 minutes or less, while for 14 m second^{-1} winds, spotting occurs in less than 7 minutes in every case. If a home is undefended during a WUI fire, these firebrands pose a serious threat to the home.

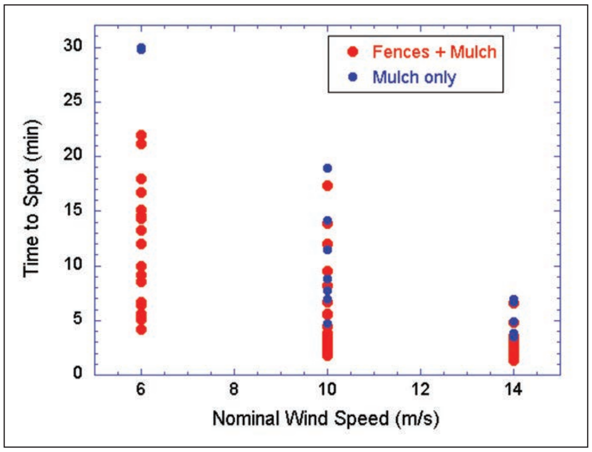


Figure 5—Spotting time as a function of nominal wind speed for experiments on mulch beds only (blue) and combinations of fence and mulch bed (red).

Firebrand spotting consists of three mechanisms: firebrand generation, firebrand transport, and ignition of the surrounding fuels (Koo et al. 2010). High speed winds break off and loft firebrands more readily. They also transport firebrands faster and farther. The ability of firebrands to ignite a spot fire in the mulch bed depends on many factors, including the characteristics of the firebrand and mulch bed, the contact between firebrand and mulch, and the local environment at the location of the firebrand. Higher speed winds deliver more oxygen to the ignition site and support smoldering (Filkov et al. 2016). However, if a critical wind speed is exceeded, the firebrand may be quenched by the cooling effect (Song et al. 2017).

Figure 6 shows that there is not a strong relationship between the time to spot and the separation distance between the end of the fence or mulch bed and the wall of the structure. This suggests that the spotting time is controlled by either firebrand generation or ignition, and that transport is not an important factor in this set of experiments, where the distance between fence and structure was relatively short.

In these experiments, spot fires appeared to be ignited by single firebrands. Not every firebrand landing in the target mulch bed found conditions favorable for ignition. Typically, a handful of spot fires (seldom more than 10) were ignited in the time period before 1 of those fires reached the wall.

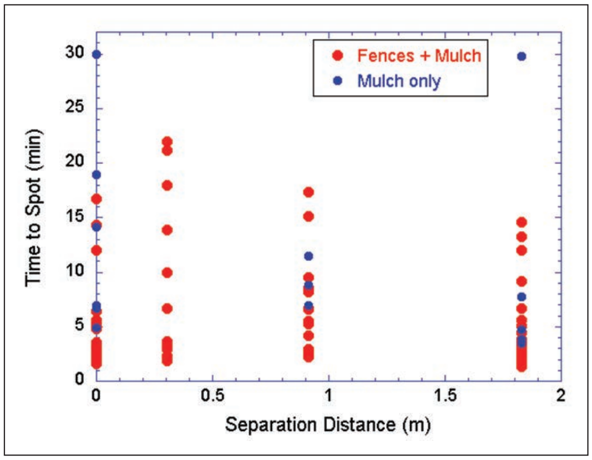


Figure 6—Spotting time as a function of separation distance for experiments on mulch beds only (blue) and combinations of fence and mulch bed (red).

Fences Without Mulch

In the absence of mulch beneath the fence, firebrand spotting was generally considerably slower, if it occurred at all; spotting occurred in only 8 of the 22 experiments with fences alone. The times to spot for these experiments are shown in figure 7.

Without mulch, the ignited fences tended to smolder rather than flame. This was a slow process that in the majority of cases did not result in spot fires. An exception is shown in figure 8, in which a smoldering piece of the fence has broken off at high wind speeds and ignited a spot fire near the wall.

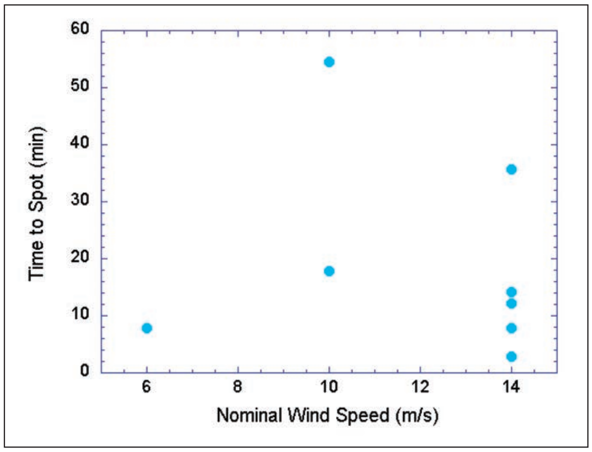


Figure 7—Spotting time as a function of nominal wind speed for experiments on fences only in which spotting occurs.

Exceptional Cases

This set of experiments demonstrated some special cases for which the fire behavior differed significantly from similar experiments.

Double Lattice Fences

A single redwood lattice fence combined with a shredded hardwood mulch bed burned with flames staying close to the ground, as shown in figure 9. This image was taken 12 minutes into the experiment. The behavior is similar to that seen with the privacy fence in figure 3.

Compare this to figure 10, in which redwood lattice fence panels have been attached to both sides of the end posts, with a spacing of 0.09 m between them. This image was taken 3 minutes into the experiment, at which point the double lattice fence is fully engulfed. The space between the fences is partially shielded from the wind field, which promotes flame attachment and spread. The changes in convective heat transfer introduced by the second fence, plus the radiative exchange between the fences, act to intensify the fire.

The time to spotting was 15 minutes in the case of the single lattice fence and 7 minutes for the double lattice fence, after the peak fire behavior in each case.

The fire behavior of the double lattice fence is sufficiently enhanced that the mulch bed beneath the fences is not necessary. Figure 11 shows a double lattice fence without mulch beneath, at 4 minutes into the experiment. Spotting in the target mulch bed occurred at 7 minutes, after the fence had collapsed into a burning pile on the ground.

Parallel Privacy Fences

The addition of a second western redcedar privacy fence parallel to the first changed the fire behavior in a similar way to the double lattice fence. Figure 12 shows a single privacy fence with hardwood mulch beneath after 20 minutes. In figure 13, a second privacy fence was arranged at a spacing of 0.20 m from the first. This could occur, for example, if two neighbors decided to build privacy fences on the property line of their respective parcels. This configuration greatly enhanced the fire behavior. Figure 13 shows the conditions 5 minutes into the experiment.

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018.

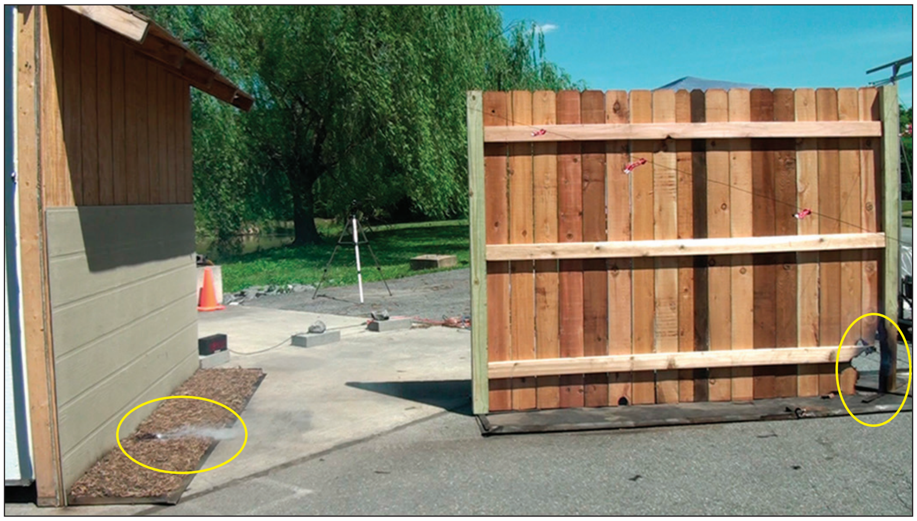


Figure 8—Firebrand spotting for western redcedar privacy fence with high wind speed (14 m second^{-1}) and at 1.8 m separation distance from wall.

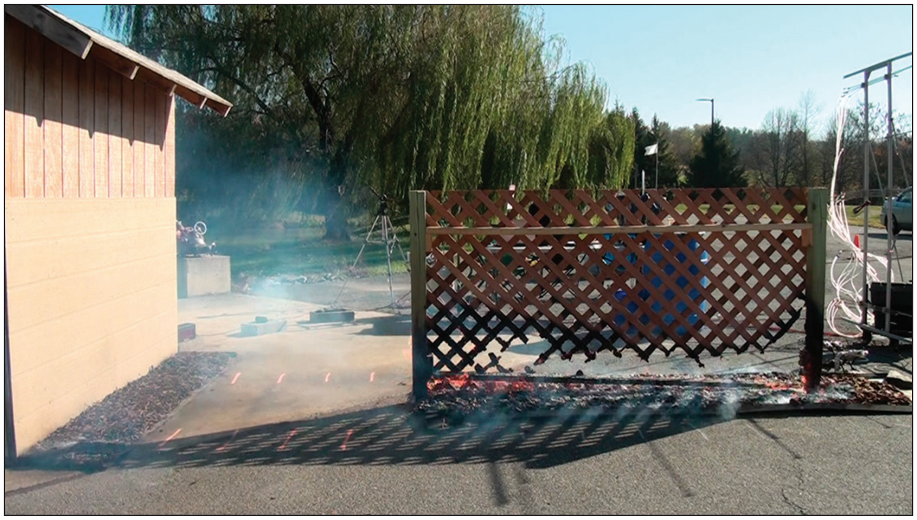


Figure 9—Single redwood lattice fence in hardwood mulch at low wind speed (6 m second^{-1}).

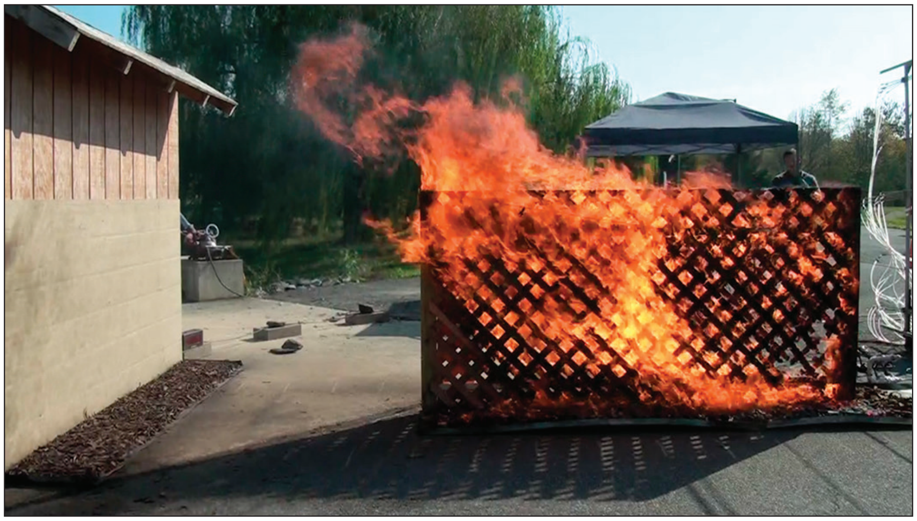


Figure 10—Double redwood lattice fence in hardwood mulch at low wind speed (6 m second^{-1}).

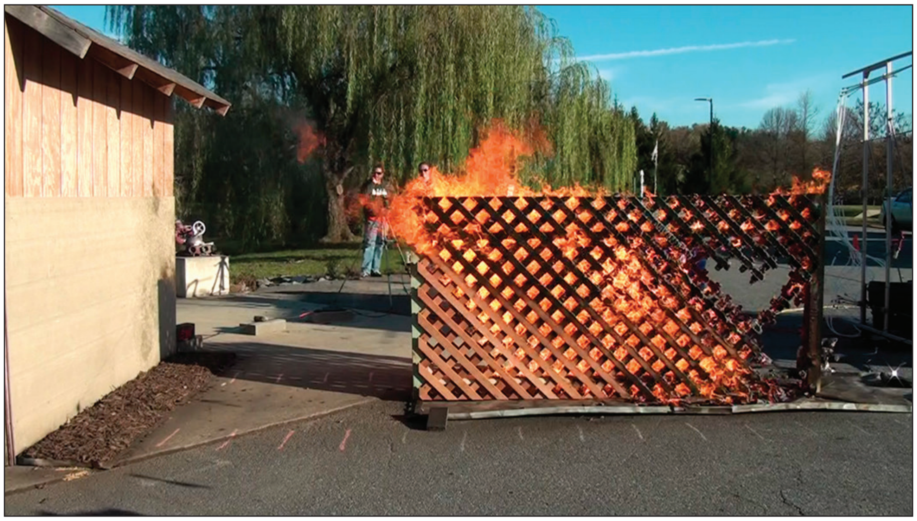


Figure 11—Double redwood lattice fence without mulch at low wind speed (6 m second^{-1}).

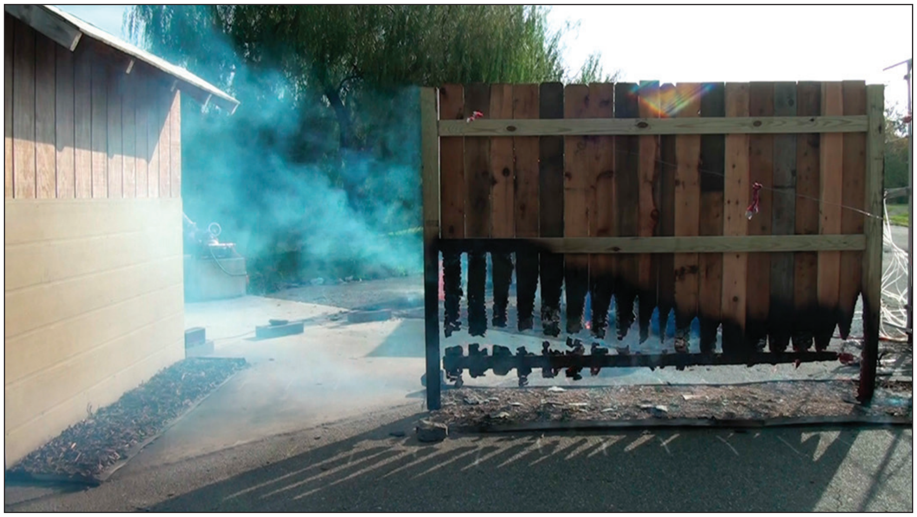


Figure 12—Single western redcedar privacy fence in hardwood mulch at low wind speed (6 m second^{-1}).

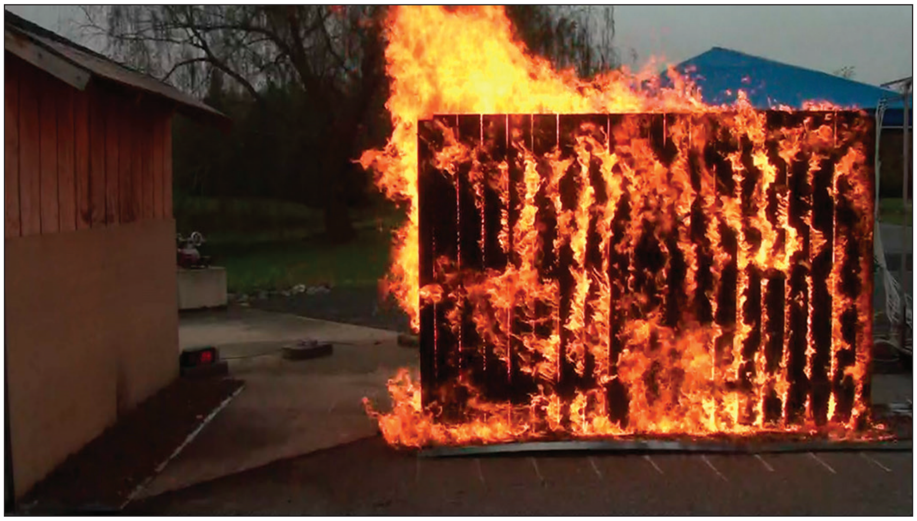


Figure 13—Parallel western redcedar privacy fences separated by 0.20 m in hardwood mulch at low wind speed (6 m second^{-1}).

Butler, Kathryn; Johnsson, Erik L.; Tang, Wei. "Structure Vulnerability to Firebrands from Fences and Mulch." Paper presented at The Fire Continuum Conference, Missoula, MT, US. May 21, 2018 - May 24, 2018.

The time to spotting was 13 minutes for the single privacy fence and 5 minutes for the parallel privacy fences.

Unlike the double lattice fences, a mulch bed beneath the parallel privacy fences was necessary for enhancing the fire behavior. Figure 14 shows that without the mulch bed the parallel fences smoldered slowly, similar to the behavior seen in figure 8 for a single privacy fence without a mulch bed beneath.

Pine Straw Mulch

A final example of unusual fire behavior was encountered with a bed of pine straw mulch. This mulch burned intensely and rapidly. However, the firebrands produced by pine straw mulch were too fine to ignite the target mulch bed. Figure 15 shows a pine straw mulch bed in direct contact with the target hardwood mulch bed at the base of the structure. Although the flames have reached the target mulch bed, no ignition took place.

If the pine straw mulch bed was combined with a western redcedar privacy fence, however, the pine straw quickly ignited the whole bottom of the fence, and spot fires were ignited in the target mulch bed by firebrands from the fence.

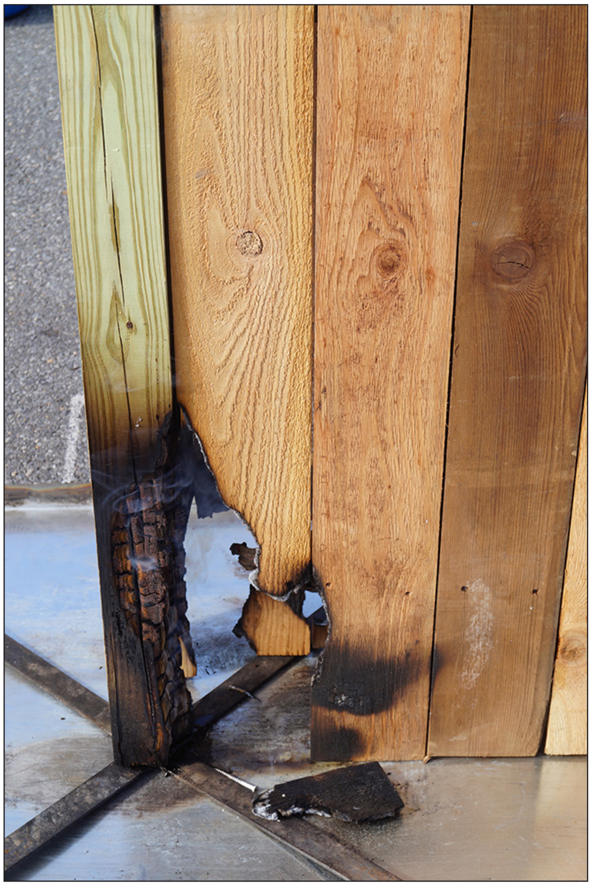
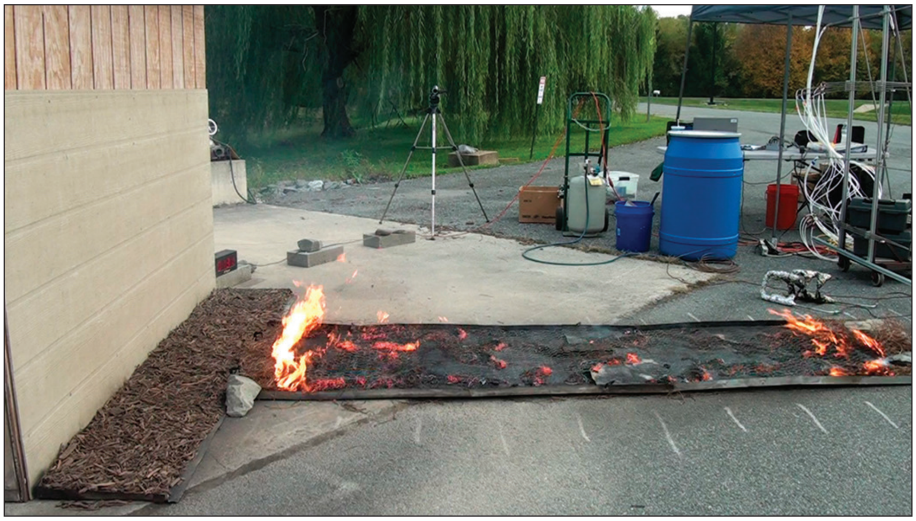


Figure 14—Parallel western redcedar privacy fences without mulch at low wind speed (6 m second^{-1}).

Figure 15—Pine straw mulch bed in contact with target hardwood mulch bed.



Removing the Structure

An additional three experiments were performed in which the small structure was removed from the area downwind of the firebrand source, the target mulch bed was moved to a distance 23 m from the source, and the burning source was subjected to a wind field of 14 m second^{-1}. The space between the source and target was asphalt and concrete, representing a worst case (i.e., favorable) scenario for transport of the firebrands over the ground. Roads and driveways make this a realistic condition for a WUI neighborhood. Figure 16 shows the experiment in which a double lattice fence has been ignited. A bed of shredded hardwood mulch and a woodpile were used in the other two long-range experiments. In each case, spot fires ignited in the target mulch bed 23 m from the firebrand source within 5 minutes after the wind machine was set to deliver high wind speeds. It should be noted that most of the spot fires occurred in the middle of the target mulch, indicating that the firebrands were lofted at some point rather than simply moving over the surface of the ground.

CONCLUSIONS

This limited series of field experiments on ignited mulch beds, fences, and combinations of fence and mulch bed in a wind field in front of a structure demonstrates that firebrand spotting may occur within 2 to 20 minutes of ignition. Spotting often occurred after peak flaming and was affected by wind fields near the structure.

For this set of wind velocities and approach angle, fence configurations, and materials, the time to spotting tended to decrease with increasing wind speed, but it did not show a strong relationship with separation distance. This is consistent with the wind having important effects on firebrand generation and on the local ignition environment.

For this series, the combination of a fence and a mulch bed appeared to decrease the time to spotting over either the mulch bed or the fence alone.

In the absence of a structure, firebrand spotting can occur within a few minutes even at long range.

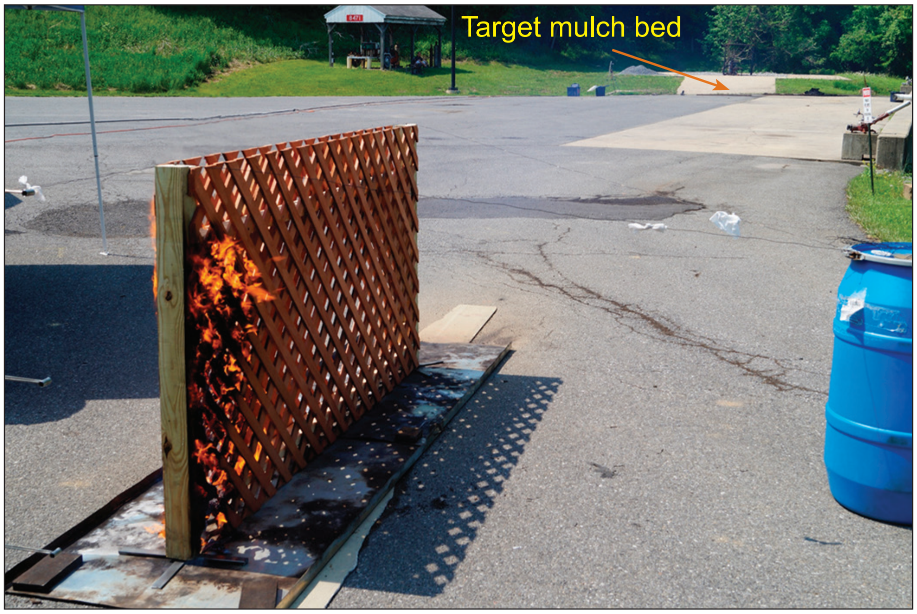


Figure 16—Double lattice fence experiment without a structure and with a mulch bed situated 23 m from the far end of the fence.

Future experiments will include effects of mitigation, including coatings and fence height above the ground, and aging on the generation and spotting of firebrands from fences.

ACKNOWLEDGMENTS

The Frederick County Fire and Rescue Training Facility in Frederick, Maryland made this work possible by allowing us to perform the experiments on its property. Discussions with Alex Maranghides have guided this work. Our thanks also go to Marco Fernandez, whose help with instrumentation, logistics, and operations has been essential to running the experiments, to Mariusz Zarzecki for his help with both experiments and analysis, and to Eric Auth, Rachel McIntyre, and Michael Pryor for their contributions as summer students.

REFERENCES

Filkov, Alexander; Kasymov, Denis; Zima, Vladislav; Matvienko, Oleg. 2016. Experimental investigation of surface litter ignition by bark firebrands. Advanced Materials in Technology and Construction; AIP Conf. Proc. 1698, 060004. https://doi.org/10.1063/1.4937859.

ICC and NARCD Councils. 2013. Communities dealing with wildland/urban interface fire; Wildfire Safe, Sound & Code Smart; WUI fact sheet. https://nanopdf.com/download/wssampcs-wui-fact-sheet-c-international-association-of-wildland-fire_pdf. [Accessed 2018 June 21].

Koo, Eunmo; Pagni, Patrick J.; Weise, David R.; Woycheese, John P. 2010. Firebrands and spotting ignition in large-scale fires. International Journal of Wildland Fire. 19: 818–843.

Maranghides, Alexander; McNamara, Derek. 2016. 2011 Wildland urban interface Amarillo Fires Report #2—Assessment of fire behavior and WUI measurement science. Technical Note NISTTN 1909. Gaithersburg, MD: National Institute of Standards and Technology. 153 p. https://doi.org/10.6028/NIST.TN.1909.

Maranghides, Alexander; McNamara, Derek; Vihnanek, Robert; [et al.]. 2015. A case study of a community affected by the Waldo Fire—Event timeline and defensive actions. Technical Note NISTTN 1910. Gaithersburg, MD: National Institute of Standards and Technology. 213 p. https://doi.org/10.6028/NIST.TN.1910.

Martinuzzi, R.; Tropea, C. 1993. The flow around surface-mounted, prismatic obstacles placed in a fully developed channel flow. Journal of Fluids Engineering. 115: 85–92.

McCaffrey, B.J.; Heskestad, G. 1976. A robust bidirectional low-velocity probe for flame and fire application. Combustion and Flame. 26: 125–127.

McGrattan, Kevin B.; McDermott, Randall J.; Weinschenk, Craig G.; [et al.]. 2013. Fire Dynamics Simulator Users Guide. Special Publ. NISTSP 1019, 6th ed. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.sp.1019.

Song, Jiayun; Huang, Xinyan; Liu, Naian; [et al.]. 2017. The wind effect on the transport and burning of firebrands. Fire Technology. 53: 1555–1568.

Taylor, Barry N.; Kuyatt, Chris E. 1994. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical Note NISTTN 1297, 1994 ed. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.tn.1297.

Evaluation of non-destructive volumetric testing methods for additively manufactured parts

Anne-Françoise Obaton[1]*, Bryan Butsch[2], Stephen McDonough[3], Ewen Carcreff[4], Nans Laroche[4], Yves Gaillard[5], Jared B. Tarr[1], Patrick Bouvet[5], Rodolfo Cruz[3], Alkan Donmez[1]

[1] Engineering Laboratory, National Institute of Standards and Technology (NIST), 100 Bureau Dr. MS 8220, Gaithersburg, MD 20899, USA

[*] Laboratoire National de Métrologie et d'Essais (LNE), 1 rue Gaston Boissier, Paris, France

[2] The Modal Shop, Inc., 3149 E. Kemper Rd, Cincinnati, Ohio 45241, USA

[3] OKOS Solutions, LLC, 7036 Tech Circle, Manassas, Virginia, USA

[4] The Phased Array Company (TPAC), 8 bis rue de la garde, 44300, Nantes, France

[5] Centre Technique des Industries de la Fonderie (CTIF), 44 avenue de la Division Leclerc, 92318 Sèvres, France

ABSTRACT

Additive manufacturing enables the production of customized and complex parts. These two aspects are very attractive for aerospace and medical sectors. However, in these critical sectors, governed by strict safety requirements, the quality of the parts is of paramount importance and the technology has advanced at a much faster pace than regulations and quality controls. The reliability of the parts must be guaranteed, and hence quality control is needed. Considering the complexity of additively manufactured part shapes, the inspection methods need to be non-destructive, three-dimensional, and volumetric. X-ray computed tomography is presently the most appropriate method, but the relative high cost and testing duration make routine inspection difficult. Thus, alternative non-destructive volumetric methods are required. In this paper, four alternative methods, utilizing acoustic waves (resonant acoustic method, process compensated resonance

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

testing) and ultrasonic waves (conventional ultrasonic testing, phased array ultrasonic testing combined with total focusing method), are investigated and compared with X-ray computed tomography using synchrotron radiation.

Keywords

Non-destructive testing (NDT), volumetric NDT, additive manufacturing (AM), acoustic methods, ultrasound methods, X-ray tomography.

Introduction

One of the main advantages of additive manufacturing (AM) is to allow the fabrication of very complex-shaped parts, including with inner cavities, that cannot be manufactured with conventional methods from a single block of material. Another main advantage of AM is the possibility to build customized parts for roughly the same price as some high-volume parts. These advantages are very much appreciated in the aerospace and medical sectors. But AM has been shown to produce specific types of defects in as-built parts, such as cross-layer defects, layer-specific defects, and unconsolidated/trapped powder defects. In these critical aerospace and medical sectors, the quality requirements of the parts are particularly severe, and the parts must be certified before they can be used in their intended application. This is a challenge for quality control of parts with such complex geometry and inner cavities, especially those with the high surface roughness levels typically of AM surfaces. The inspection methods must be non-destructive, volumetric to screen inner cavities, appropriate for high roughness and complex shapes, but also inexpensive and fast for routine inspections. Among existing volumetric non-destructive testing (NDT) methods, X-ray tomography is presently the most appropriate method to inspect the inner features and defects of complex AM parts. However, X-ray tomography is expensive and the scanning process takes a long time. In addition, the resulting scan files are very large; thus, they are not easily viewable. So, even if X-ray tomography is recommended for first article inspection (FAI), it is not suitable for routine inspection of high-volume production of customized parts, for example. Alternative methods to X-ray tomography must be explored [1, 2].

First, all the NDT volumetric methods that are currently used for the characterization of conventionally manufactured parts require a new understanding for how they interact with AM parts. Then, further development of existing NDT methods, or novel ones, are greatly needed to handle the level of complexity that AM presents.

The resonant acoustic and ultrasound NDT linear volumetric methods are potentially valuable techniques that need to be explored for AM parts.

This paper presents the capabilities of some of the existing acoustic and ultrasonic volumetric NDT methods to inspect AM parts with specific defects. The artifacts[1] proposed by ISO TC261/ASTM F42 Joint Group 59 (JG59) "NDT for AM parts" are used in this study. The paper also presents results obtained with X-ray computed tomography (XCT) using synchrotron radiation as a comparison.

The paper is structured as follows: the first section is dedicated to the presentation of the JG59 and their standards development regarding NDT inspections using a specially designed artifact; the second section describes the fabrication of this artifact in different materials; and the last section presents the principles, inspection results on the artifacts, and benefits/drawbacks of the following five volumetric NDT methods:

- whole-body inspection methods using acoustic waves – (1) resonant acoustic

[1] The English spelling of “artefact” will be used instead of the American spelling of “artifact” in the ISO/ASTM standard.

method (RAM) and (2) process compensated resonance testing (PCRT),

- selective inspection methods using ultrasonic waves – (3) conventional ultrasonic testing (CUT) and (4) phased array ultrasonic testing (PAUT) methods, and
- X-ray for comparison – (5) X-ray computed tomography (XCT) using synchrotron radiation (SRXCT).

STANDARDIZATION GROUP ISO TC261-ASTM F42 ON AM – JOINT GROUP JG59 ON NDT FOR AM PARTS

Since AM is causing a revolution in manufacturing capabilities, many industries are interested to take advantage of these methods. However, each time a new technology grows in capability and use, standardization issues arise that must be overcome for the technology to continue its advancement. Several standards developing organizations (SDOs) are addressing this subject. The International Organization for Standardization (ISO) and ASTM International have implemented a formal agreement to co-develop a series of AM standards through several joint groups (JG). JG 59 of ISO Technical Committee (TC) 261 and ASTM F42 is dedicated to NDT for AM parts.

Goal of the JG59

Several standards dedicated to NDT of manufactured parts already exist, and the JG59 does not aim to duplicate these standards. AM is very different from other manufacturing processes like casting, forging, machining, etc. With AM, parts are built layer by layer from raw feedstock such as metal powder, so one might say that the part is manufactured simultaneously with the material. This also means that the microstructure of the built material and the potential for embedded defects are specific to the AM technology. As mentioned previously, AM enables manufacturers to build very complex geometries, though with rough surfaces. Therefore, existing NDT methods must be evaluated for AM applicability. These methods might need to be adapted to accommodate AM characteristics. Furthermore, new NDT methods may also be needed. Thus, a primary goal of the JG59 "NDT for AM parts" is to write a best practice guide presenting potential NDT methods suitable for post-process inspection of AM metallic parts, referring to existing NDT standards when appropriate. This best practice guide will contain a list of typical AM defects that can be found in two of the main types of metal additive manufacturing process categories used in industry: power bed fusion (PBF) and directed energy deposition (DED). It will give recommendations for NDT methods suitable for the special features of AM. A first version of this guide should be complete by the end of 2018.

To develop this guide, joint group members have designed and fabricated several artifacts with typical AM defects in different materials. These artifacts were then characterized with different NDT methods by the JG59 members (Table 1) to evaluate the behaviors and potential for each NDT method for use with AM parts.

TABLE 1 NDT technologies investigated by JG59 with star artifacts in different materials.

Material	NDT Technologies	Center/Company/University
Hastelloy X	X-Ray Computed Tomography (XCT)	MTC, UK
	Phased Array Ultrasonic Testing (PAUT)	University of Bristol, UK
	Process Compensated Resonance Testing (PCRT)	Vibrant, Germany
	Nonlinear acoustic method (NLA)	Theta Tech, UK
	Thermography testing (TT)	University of Bath, UK
	Digital Radiography testing (RT)	FujiFilm, USA
Maraging Steel	XCT	MTC, UK
	PAUT	University of Bristol, UK
	PCRT	Vibrant, Germany
	NLA	Theta Tech, UK
	TT	University of Bath, UK
	RT	FujiFilm, USA
Cobalt Chrome	Resonance Acoustic Method (RAM)	NIST/LNE & The Modal Shop, USA/France
	XCT	NIST/LNE, USA/France
	Synchrotron XCT (SRXCT)	Novitom, CTIF, LNE, France & NIST, USA
Stainless Steel	XCT	EWI, USA
	PAUT	EWI, USA
	XCT-Low Energy	ISS, Germany
	XCT	MTC, UK
	RAM	NIST/LNE & The Modal Shop, USA/France
	PCRT	NIST/LNE & Vibrant, USA/France
Aluminum	XCT	MTC, UK
	Conventional Ultrasonic Testing (CUT)	NIST/LNE & OKOS, USA/France
	PAUT- Total Focusing Method (TFM)	TPAC, LNE, France & AOS, NIST, USA
Titanium	XCT	MTC, UK
Titanium	SRXCT & Neutron diffraction ND	ESRF & ILL, France

Star artifact with typical AM defects

This artifact was designed by JG59 to be optimized for laboratory 450 kV X-ray computed tomography (XCT). The size of the artifact depends on the density of the material. It is in the shape

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

of a star (Fig. 1), hence its name: star artifact. The artifact contains the following defects unique to AM parts:

1. Cross-layer defects that are represented by vertical cylinders of different diameters but the same length. These cylinders are connected to each other and open to outside at the bottom of the star, so that powder is released at the largest diameter cylinder;
2. Layer-specific defects that are represented by horizontal cylinders of different diameters but the same length, with an open end to release powder;
3. Unconsolidated/trapped powder defects that are represented by spheres of different diameters and internal cylinders in various orientations.

Voids and porosities are common defects found also in conventional manufacturing and are mostly covered by existing NDT standards.

These defects are located into critical areas such as deep sections and hard-to-reach areas, in five different regions. Two versions of the star design, designated as S1 and S2, were used in the evaluation of NDT methods (Fig. 1). In these two versions the defects are of the same size, same height along the part, and the same orientation; the only difference is their locations. The S2 design has the defects in thinner sections of the star branches, while the S1 design has the defects in the thicker sections. All defects are in the range of 100 µm-800 µm in diameter.

At NIST, we evaluated statistical NDT methods that required a large number of samples. Considering this fact, we decided to build half-size stars also. However, the sizes of the defects remained the same.

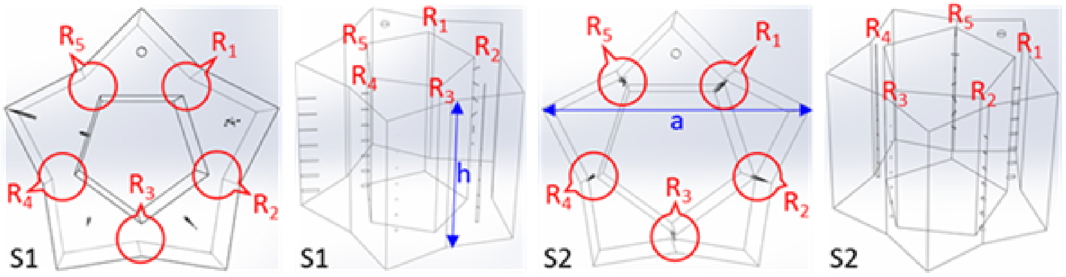


FIG. 1 Schematics of the AM star artifacts (design S1 and S2) proposed by JG59 (where R is a Region in the artifact, h and a define the height and width of the artifact, respectively, and numerical values are given in Table 2).

STAR ARTIFACT FABRICATION

The star artifacts were manufactured by different members of the JG59, depending on their capabilities in terms of AM equipment and materials. NIST built star artifacts in cobalt chromium (Co-Cr) (Figs. 2 a and b) and in stainless steel (SS) (Figs. 2 c and d) using AM machines belonging to the laser PBF process category. Zodiac Aerospace built aluminum star artifacts using also a laser PBF process. The characteristics of the manufactured star artifacts are shown in Table 2 and a picture representing the finished star artifacts is shown in Fig. 3. Since some of the NDT methods investigated require statistical analysis, significantly more artifacts (specifically, SS artifacts) were built for those methods to establish the baseline for artifacts with no defects. Furthermore, in order to evaluate the capability of such techniques to sort parts based on the numbers of defects they contain, stars with different numbers of defects were manufactured in two very similar builds (i.e., the same location on the AM build platform, with two or three marginally different parts added on the platform between the first and second build):

- ➢ 8 similar parts with defects in the 5 Regions, such as defined by the JG59;

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

- 8 similar parts with defects in only Region 1;
- 8 similar parts with defects in only Regions 1 and 2;
- 8 similar parts with defects in only Regions 1, 2, and 3;
- 8 similar parts with defects in only Regions 1, 2, 3, and 4.

Four of the eight stars are from one build, the other four are from a second build.

TABLE 2 Characteristics of the manufactured star artifacts (S0 refers to a fully dense part, without inner features representing the defects, "a" refers to width of the star, and "h" to its height, see Fig. 1)

Material	Version	Size: full /half (HF)	Quantity
Cobalt-chrome-molybdenum-based superalloy CoCr MP1	S0	Full (a=60 mm, h=45 mm)	1
	S2		3
	S0	Half (size of defects/2 also (a=30 mm, h=22.5 mm)	1
	S2		1
Aluminum AlSi10Mg	S1	Full (a=114 mm, h=45 mm)	1
	S2		1
Stainless steel 17-4	S0	Full (a=60 mm, h=45 mm)	2
	S1		2
	S2		2
	S0	Half (a=30 mm, h=22.5 mm)	40
	S1		8
	S2		8
	S2	HF without defect in Region 1	8
	S2	HF without defect in Regions 1 and 2	8
	S2	HF without defect in Regions 1, 2 and 3	8
	S2	HF without defect in Regions 1, 2, 3 and 4	8

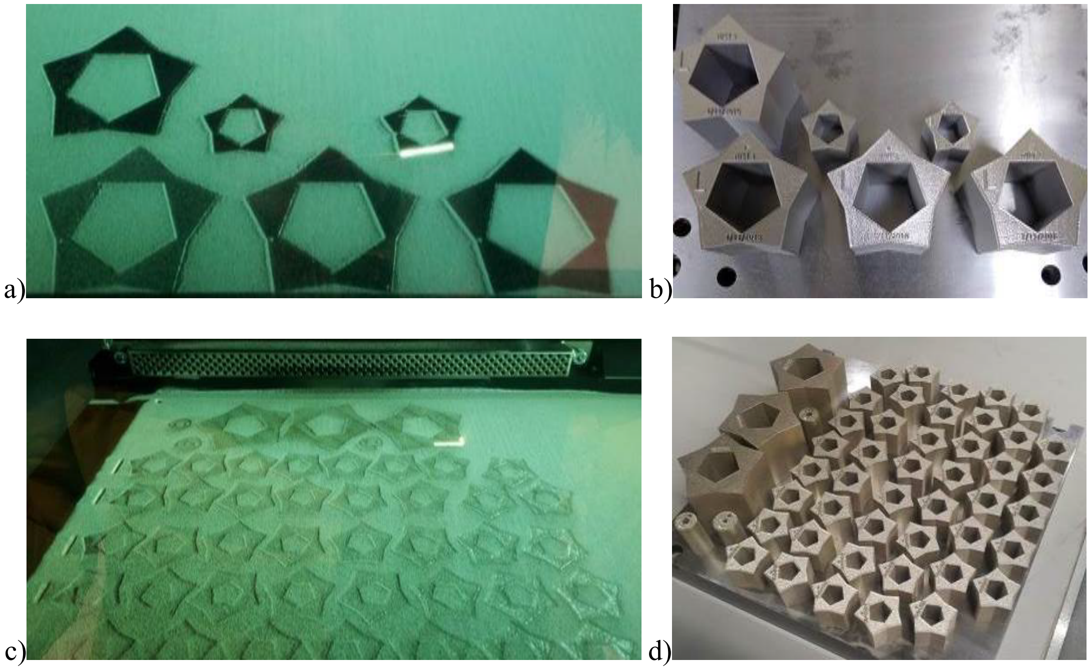


FIG. 2 Photos of the manufacturing of the Co-Cr (a and b) and SS (c and d) star artifacts, during the build (a and c) and after the build on the machine platform (b and d).

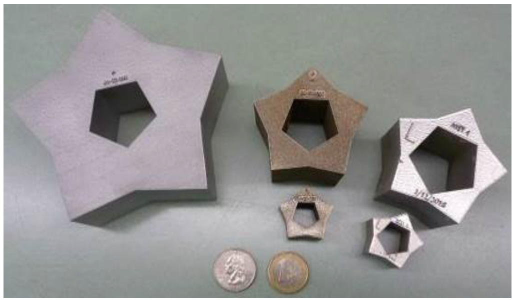


FIG. 3 Photos of the manufactured star artifacts, from left to right: Al, SS, and Co-Cr.

VOLUMETRIC NDT METHODS EVALUATED

Non-destructive inspection methods can be separated into two main groups: (1) surface methods enabling only a surface inspection of the test object (e.g., by visual, dye penetrant,

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

magnetic particles, etc.) and (2) volumetric methods enabling volumetric inspection of the test object (e.g., by eddy current, ultrasound, radiography, thermography, acoustic emission, X-ray tomography, etc.) [3]. Volumetric methods can also be divided into two main groups: whole-body inspection methods and selective inspection methods. The whole-body inspection methods do not provide information about defect locations, whereas selective inspection methods do provide such information. We have investigated volumetric methods from both groups in order to evaluate their respective potential for inspecting AM parts.

Whole-body inspection methods

Resonant ultrasound spectroscopy (RUS) methods

Principle of resonant ultrasound spectroscopy (RUS) methods

There are several variants of resonant ultrasound spectroscopy (RUS) methods, as described in ASTM E2001 [4], but their basic principles are similar. These whole-body inspection approaches compare the frequency spectrum of the mechanical resonances of a set of reference parts, known as “good parts”, with the frequency spectrum of the mechanical resonances of the test parts. These are pass/fail assessments. Two similar objects will have similar resonant frequency spectra. A shift of the frequency peaks of a part, compared to the statistical variation from the similar good parts, will be the signature of a structural change of the part (e.g., changes in part geometry and/or material properties associated with the mass, stiffness, and damping).

A characterization using RUS methods includes several steps. The first step consists of a mechanical impulse of the test object to make it vibrate as free as possible. Under this excitation, the object vibrates at its natural resonant frequencies. At the second step, the response of the test object is monitored by a sensor, and the frequency spectrum is recorded. At the third step, this spectrum is compared to the spectrum of the good parts, which are obtained following the same first two steps. For this comparison, first the well-defined resonant peaks for each good part are identified. A subset of such resonant peaks that are consistent for all reference parts peaks and have distinct separation with bad parts are selected for further consideration. The ranges (boundaries) in variations in each resonant peak frequency of the good parts are evaluated, these are called “criteria”. Outliers of these variations must be excluded from these ranges. The measured resonant peak frequencies of the test parts are compared against the corresponding ranges in frequencies of the good parts. All test parts with frequency peaks outside of these boundaries are rejected as faulty parts.

Resonant acoustic methods (RAM) and process compensated resonance testing (PCRT), described in ASTM E2534 [5], are two types of RUS methods.

Principle of resonant acoustic method (RAM)

In the resonant acoustic method (RAM), the excitation is done by an impact hammer (first step) and the resulting acoustic frequency data is obtained with a microphone. There are no contact probes that need to interact with the test part. This allows the parts to resonate in as free a state as possible. The output of the microphone is converted into a frequency spectrum using a high-speed analog to digital converter (24 bits) performing a fast Fourier transform to determine resonant peaks (second step). Finally, a statistical analysis conducted with software reveals if the test parts fall within the defined criteria or not (third step). If at least one of the measured resonant peak frequencies of the test part lays outside of the established frequency ranges of the good part, the test part is considered bad.

Resonant acoustic method (RAM) inspection of the SS half-size star artifacts

Fig. 4 shows the setup used for the resonant acoustic method (RAM) measurements [6] of eighty-eight SS half-size stars (Fig. 5), resulting from the two successive builds (Table 2). Among them, twenty from each build were manufactured as reference parts with no defects and were considered as reference parts to define the baseline for the criteria. The stars were systematically weighed before the test as the inspection software enables weight compensation. Then, the star artifacts were impacted with an automatic hammer on the sides of their bottom faces, as these positions were experimentally found to provide the least amount of damping. Each part was impacted three consecutive times at three different branches of the star at the base. This procedure allowed us to check for repeatability and reproducibility. For consistent impacts on every part, the same positions on the same branches of the stars were always struck. The frequency window considered for analysis ranged from 500 Hz to 50 kHz, and a resolution of 3.5 Hz was chosen which gives an acquisition time of 0.23 s.

We considered the first and second builds separately. The data on the twenty reference parts were first gathered, and then the rest of the twenty-four stars with defects were tested. The tests were repeated on two consecutive days. The acceptance criteria to define the baseline for comparison of the resonant peaks, from this first set of data, were set up over six frequency bands, as shown in. Fig. 6. This figure shows the complete spectrum for a single part test in blue. The vertical green lines show the bands of the resonant peaks chosen for evaluation. These criteria are used by the software to sort the bad parts from the good. The data for a part must pass all the criteria in order to pass resonance testing.

There is a constant shift, for all peaks, in the frequency spectrum between build 1 and build 2 (Table 3). However, the same criteria (same width) could be chosen between build 1 and build

2. It is common (especially in a powder metal build process) to see small shifts in resonant spectra from build to build due to variations in the raw material, manufacturing conditions, environmental conditions, etc. It should be noted that the general shape of the frequency spectrum does not change from build to build, just a constant shift in frequency is observed, based on changes to the mass or stiffness of the parts. It should also be noted that the spectrum shift is a global shift for the entire build and does not affect the ability to sort the good parts from the defective parts.

The overall outcome of the resonant acoustic test for the eighty-eight SS half-sized star artifacts is displayed in Table 4.

We were able to define criteria that allowed us to sort the good parts from the ones with internal defects. The reference parts showed a significant difference in resonance. However, in the range of frequencies analyzed, the parts could not be sorted based on the number of defects they contain.

The main benefits and drawbacks of the resonant acoustic method (RAM) are summarized below.

Benefits of the method	➢ easy to use ➢ fast ➢ simpler and faster than PCRT ➢ no restriction in size ➢ no restriction in shape ➢ no roughness restriction ➢ no part set-up required ➢ suitable for medium to high volume ➢ able to identify defective parts ➢ suitable for production end of line testing, or routine quality inspection
Drawbacks of the method	➢ not possible to identify the type or location of defects in the part

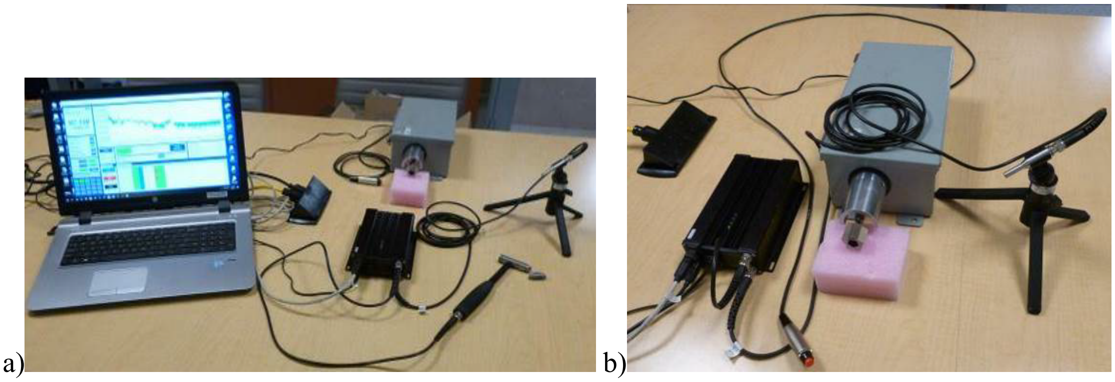


FIG. 4 The Modal Shop RAM set-up (a) and inspection of the half-size SS star artifacts using the set-up with the automatic hammer (b).

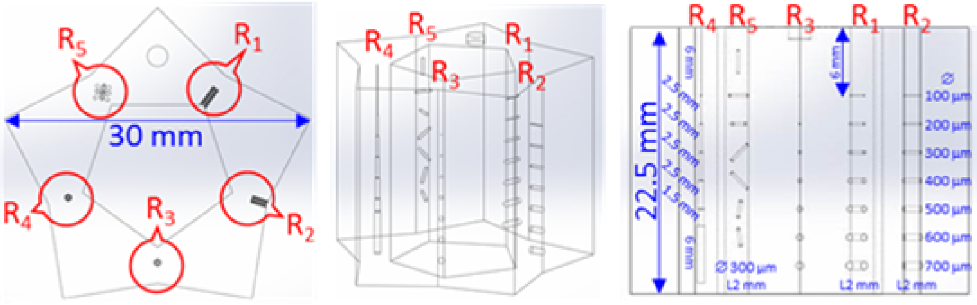


FIG. 5 Computer Aided Design (CAD) of the SS half-size star artifact, S2 design, seen from different sides ("R" refers to the Region in the star).

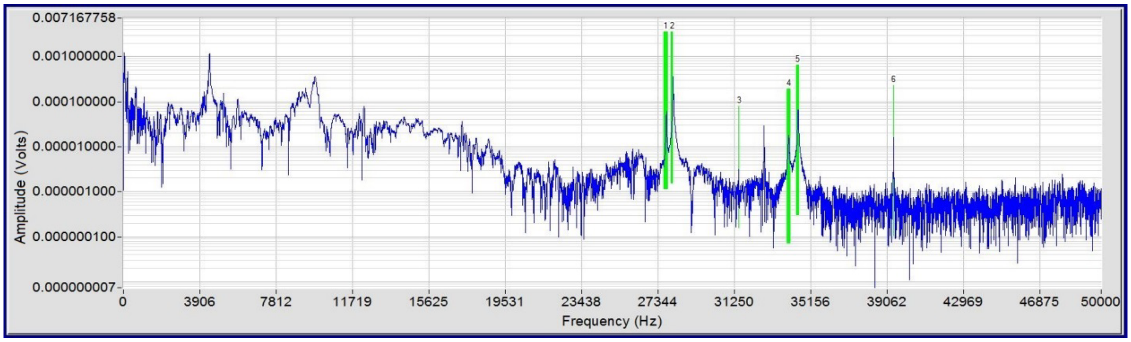


FIG. 6 Typical half-size SS star artifact RAM spectrum showing the subset of the six resonant peaks used for comparison criteria.

TABLE 3 RAM criteria, center frequencies, and shift between build 1 and build 2 for the half-size SS star artifacts.

	Center Frequency		Frequency Shift (Hz)
	Build 1 (Hz)	Build 2 (Hz)	
criterium 1	27742.5	27549.5	-193
criterium 2	28072.5	28017.5	-55
criterium 3	31489.5	31455	-34.5
criterium 4	34027.5	34001	-26.5
criterium 5	34489	34601	112
criterium 6	39401	39366.5	-34.5

TABLE 4 RAM overall Pass/Fail results for the 88 half-size SS star artifacts.

	Build 1		Build 2	
	Pass	Fail	Pass	Fail
Good parts	20	0	19	1
Bad parts with defects	0	24	0	24

Process compensated resonance testing (PCRT)

Principle of process compensated resonance testing (PCRT)

Process compensated resonance testing (PCRT) [5] uses piezoelectric transducers in contact with the part, but otherwise is very similar to RAM. Excitation of the part uses a piezoelectric transducer to generate a low-energy swept-sine wave, generally with ultrasonic frequencies (5 kHz-500 kHz). Then the frequency response is recorded with other piezoelectric transducers. Each test requires configuring a specific fixture and optimizing data collection settings for each new part geometry.

The PCRT system used to characterize the star artifacts is fully automated with one transmitter and two receiver transducers. For data analysis, the system uses a Z-score [7] to identify outliers of resonance peak variations within a population of parts. In addition, when a simple resonance analysis is insufficient due to interlacing of the spectra for good and bad parts, the system uses a pattern recognition tool such as the Mahalanobis-Taguchi system (MTS) [8] for sorting of the parts.

Process compensated resonance testing (PCRT) inspection of the SS half-size star artifacts

Fig. 7 shows the setup used for the PCRT inspection of seventy-four of the eighty-eight SS half-size stars resulting from the two successive builds (Table2). Twenty and eighteen parts were considered as reference parts from builds 1 and 2, respectively. For each part, data were collected between 25 kHz and 225 kHz in less than three minutes, and then processed using Z-score and MTS analysis.

The repeatability of resonance measurements was evaluated using one reference part, over thirty repetitions. The variations, expressed as the standard deviations, in each measured peak frequency over thirty repetitions were calculated. The frequency peaks with large variations were eliminated from the set of peaks to be used for the defect detection criteria. The standard deviations for the remaining twenty-six resonant peaks were between 0.05 % and 0.013 % of the mean peak frequencies.

PCRT partial spectra for four good and four bad parts are compared in Fig. 8, which shows that the shift in frequency between good and bad parts is not the same for every resonant peak. To deal with this, a proprietary pattern recognition algorithm (VIPR [9] for Vibrational Pattern Recognition) is used to differentiate the parts with defects from the parts with no defects. To confirm repeatability, these tests were run three times. Each test run gave the same results. The overall results are presented in Fig. 9 and in Table 5. As with RAM, we were able to sort the good parts from the ones with internal defects. However, the parts could not be sorted by the number of defects they contain.

The main benefits and drawbacks of process compensated resonance testing (PCRT) are summarized below.

Benefits of the method	➢ fast but not as fast as RAM ➢ no restriction in size ➢ no restriction in shape ➢ no roughness restriction ➢ suitable for medium to high volume ➢ able to identify defective parts ➢ suitable for production end of line testing, or routine quality inspection
Drawbacks of the method	➢ part set-up required ➢ not possible to identify the type or location of defects in the part

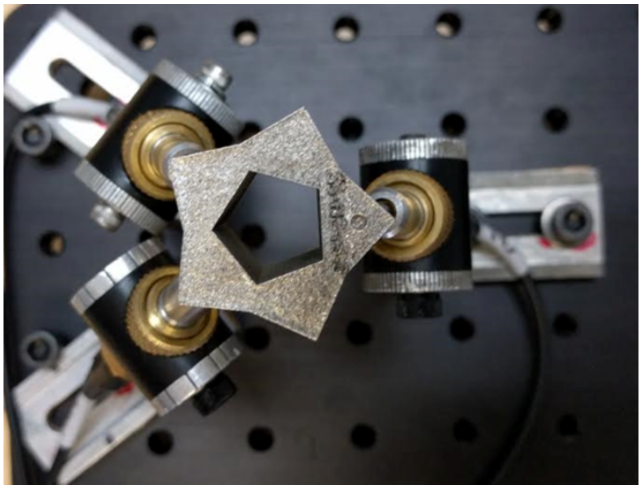


FIG. 7 PCRT set-up used for inspection of the half-size SS star artifacts.

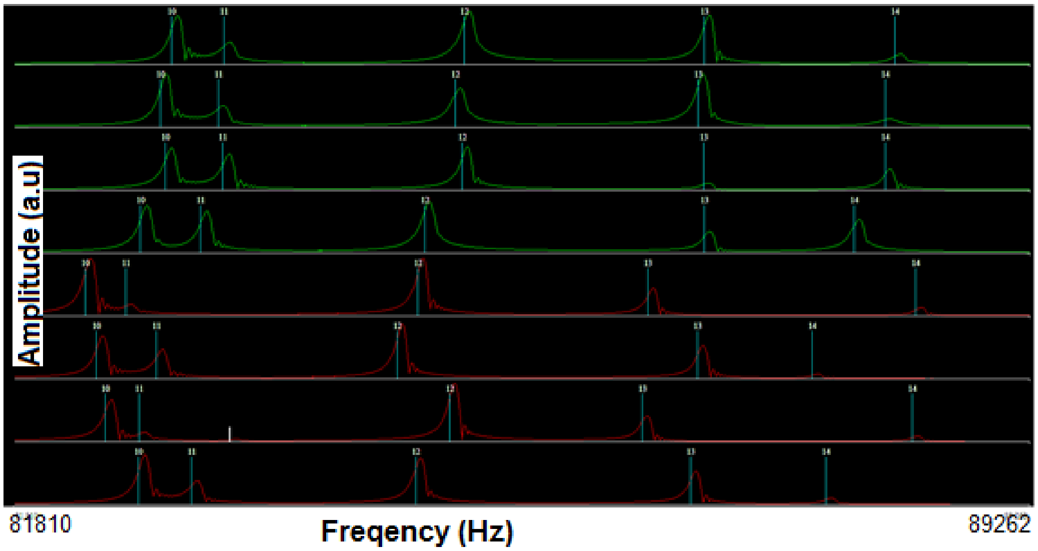


FIG. 8 Typical half-size SS star artifact PCRT spectrum. The green plots represent four good parts and the red ones represent four bad parts.

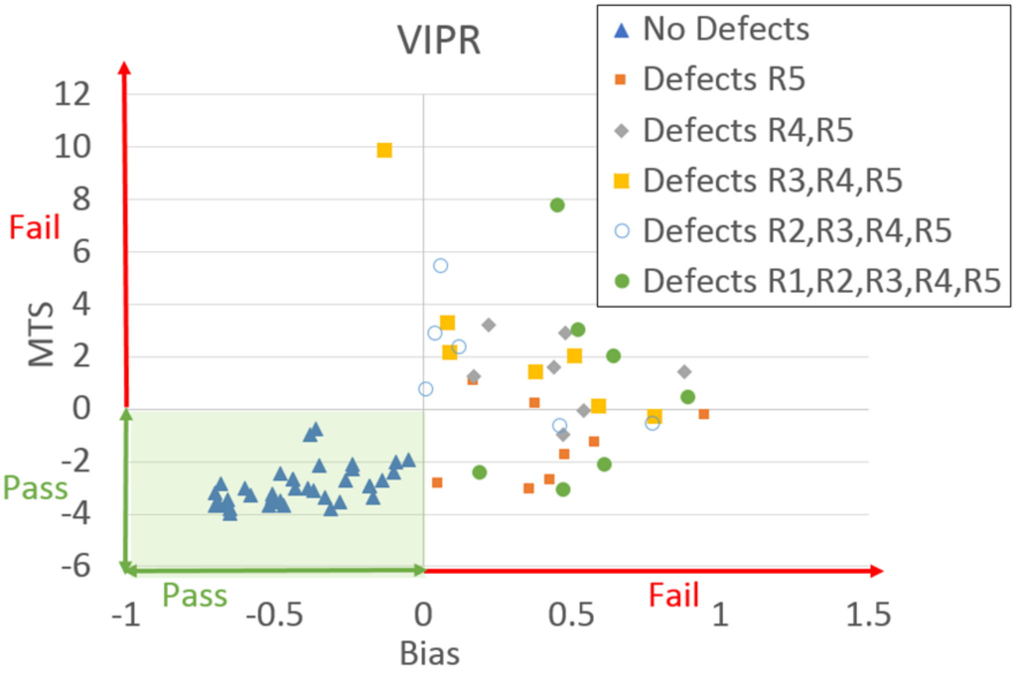


FIG. 9 PCRT overall Pass/Fail results for the 74 half-size SS star artifacts based on MTS and Bias scores.

TABLE 5 PCRT overall Pass/Fail results for the 74 half-size SS star artifacts.

	Build 1		Build 2	
	Pass	Fail	Pass	Fail
Good parts	18	0	20	0
Bad parts with defects	0	24	0	12

Selective inspection methods

Selective inspection methods use scanning techniques to inspect only selected regions of a part, possibly from multiple directions.

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

Conventional ultrasonic testing (CUT)

Principle of conventional ultrasonic testing (CUT)

Conventional ultrasonic testing (CUT), as described in ASTM E1001 [10], uses generally one or more piezoelectric transducers, which can act as transmitter, receiver, or both, to produce ultrasound into the test part. The emitting frequencies are theoretically ranging from 20 kHz upwards to 1 GHz, depending on the density of the material and on the size of the smallest defect to be detected. The ultrasound wavelength ($\lambda=v/f$, where v is the velocity of sound in the material of the tested part and f is the sound frequency) must be roughly the size of the smallest defect. Decreasing the wavelength allows detection of smaller defects but reduces the penetration depth because attenuation increases significantly. When the CUT is performed in contact mode, a thin film of couplant need to be spread on the surface of the transducer to increase the transmission of the ultrasonic waves into the part. This eliminates variations due to microscopic air pockets, which greatly impede ultrasound. Acoustic impedance of the couplant should be similar to that of the test part. But, CUT can be also performed in non-contact mode. In that case, the transducer and part are immersed in a tank of liquid (generally water) which also serves as the couplant. This allows the ultrasound beam to be scanned over various regions of a part with minimal variations in coupling.

When sound waves transmitted into the part, reflection/transmission/refraction occurs at each boundary separating a defect from the surrounding material, and at the boundaries defining the part itself. The resultant echoes are detected by a transducer and processed to determine the presence of the defects and their approximate locations. The detailed scanning procedure and signal processing must compensate for confounding phenomena such as mode conversion, scattering, and shadowing. These phenomena can prevent observation of some defects.

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

There are three possible ways of presenting the information (Fig. 10):

1. A-scan (X-axis: *time*, Y-axis: *reflected wave amplitude*): shows peaks on a graph (oscilloscope). The graph represents the amplitude of the reflected sound wave as a function of time. Position on the X-axis reflects the depth of each reflecting feature or flaw. Position on the Y-axis reflects the amplitude of the echo from each reflecting feature or flaw.
2. B-scan (X-axis: *linear position of the transducer*, Y-axis: *depth of the reflector,* grey levels or colors: *reflected wave amplitude*): shows an image with different grey levels or colors corresponding the amplitude of the reflection from each reflecting feature or flaw. Position on the X-axis shows the position of the transducer along its scan path. Position on the Y-axis reflects the depth of each reflecting feature or flaw.
3. C-scan (X-axis: *linear position of the transducer on the X axis*, Y-axis: *linear position of the transducer on the Y-axis,* grey levels or colors: *reflected wave amplitude, time of flight or depth*): shows an image viewed from the top (planar view) of a region of interest within the volume of the test specimen. The images display different grey levels or colors corresponding to the amplitude, time of flight, or depth of the signal for different positions of the transducer scanning the surface of the part. A C-scan image is formed in a plane normal to a B-scan image.

The CUT system used to inspect the star artifacts is associated with a software system (ODIS from OKOS), including amplitude, time domain and frequency domain, and imaging in real-time or post collection review. The software enables simultaneous A-B-C-scan collection. The color code on the images indicates the amplitude of the signal (i.e., change in attenuation) such as displayed in Fig. 11, where blue corresponds to low amplitude and white to high amplitude.

Conventional ultrasonic testing (CUT) inspection of the aluminum S1 star artifact

The aluminum S2 design was tested, but defects could not be detected. Therefore, only the scanning of the aluminum S1 design will be described. The computer-aided design (CAD) of the S1 aluminum star artifact is shown in Fig. 12. This part was tested in the configuration shown in Fig. 13. One transducer, acting as a transmitter/receiver, was used with the artifact immersed in water. The part was first scanned from the top, and then from each side of its branches (Fig. 13). The schematics of the scanning strategies are shown in Fig. 14. The characteristics of the transducer used for both scans are displayed in Table 6. The respective water path distances between the transducer and the test artifact during the scan were 132 mm for the top surface scan, and 83 mm for the side surface scans. The gain was adjusted to 39 dB to increase the sensitivity. Then, in order to outline the important features inside the artifact, two time-domain windows (data 1 and data 2) were defined on the A-scan, such as the one shown in Fig. 15. These windows correspond to two different Regions (depths) of interest in the artifact. Their widths (with corresponding ranges of depths) were determined based on the speed of sound in the material. The speed of sound given in the literature for wrought aluminum was chosen (as 6375 m/s). However, this value might be different in layer-by-layer manufactured aluminum. Consequently, the uncertainty on the corresponding depths of interest is not negligible. Data 1, the red rectangle on Fig. 15, excludes the top and rear surfaces of the part, whereas data 2, the green rectangle on Fig. 15, selects the Region around the rear surfaces. The C-scan images shown in Fig. 16 through Fig. 20 are based on these two time-domain windows focusing on two different Regions of depth.

Since all the defects in the artifact are in the same position along its height, the first defect does not allow the rest of the defects to be inspected by CUT from the top surface of the star. In contrast, the C-scan images recorded from the side surfaces of the star artifact and presented in Fig. 16 through Fig. 20 reveal more information:

- The cylinders with different orientations (d_{12}, between Regions 1 and 2) cannot be observed in near field from sides labeled 1A and 2B on Fig. 16, probably because the gain was constant with distance. However, the larger ones can be seen in far field from the side labeled 3B (Fig. 17), but not from the side labeled 5A (Fig. 20).
- The vertical cylinders with different diameters (d_{23}, between Regions 2 and 3) cannot be seen from the side labeled 1A (Fig. 16) because the cylinders with different orientations (d_{12}) hide them. The larger cylinders can be seen from all sides, in near and far fields.
- The spheres of different diameters (d_{34}, between Region 3 and 4) cannot be seen from any sides because the sphere shape is complex to detect with ultrasound waves. Indeed, this geometry scatters the waves in every direction. Thus, the reflected wave reaching the transducer is very low in energy. In this experiment, the energy of the reflected wave might be too low to be detected.
- The horizontal cylinders of different diameters located on an outside edge of the star (d_{45out}, between Regions 4 and 5) are not seen from most sides. This is probably due to the fact that they are positioned right at the borders of the time windows. So, since the widths of these windows are not accurate, they might be outside. However, they can all be seen and discretized in far field from the side labeled 3A (Fig. 18). They appear right in the middle of the image rather than on an edge probably because of diffraction at the R4 edge of the star.

- The horizontal cylinders of different diameters located on an inside edge of the star (d_{45in}, between Regions 4 and 5) are not seen from most sides probably because they are very small (2 mm long) and located on an edge. However, the larger ones can be seen from the side labeled A4 (Fig. 19) in near field.
- There was a mistake in the CAD data used to manufacture the aluminum S1 star artifacts. One of the cylinders with a different orientation is isolated from the others (d_5 in Region 5, Fig. 12). We might see it from the side labeled 1B (Fig. 20) in near field. The reason why it is not seen from the other sides is probably due to its orientation.

Except for the spheres, all other types of larger defects were detected from at least one side, though they were not discretized in size. However, the interpretation of the images is not straight forward and needs experience. In addition, the defects and their localization were known in advance. In the case of blind inspection, the analysis would have been even more difficult. Furthermore, dimensional measurements were not possible. However, despite the surface roughness of the AM parts, we were able to scan the star artifacts from as-built surfaces, even from side surfaces which are generally rougher than the top surfaces in AM. So, the roughness of the surface did not preclude CUT. Moreover, it should be noted that the size and thickness of the aluminum star does not allow inspection with laboratory 225 kV X-ray computed tomography (XCT). Laboratory 450 kV XCT is required. This underlines an advantage of UT methods for large parts – 450 kV XCT systems are relatively rare.

The main benefits and drawbacks of the conventional ultrasonic testing (CUT) method are summarized below.

Benefits of the method	➢ few roughness restrictions ➢ real-time imaging

	➢ less restriction in size than X-ray computed tomography ➢ able to identify the type and location of defects in the part
Drawbacks of the method	➢ part set-up required ➢ need experience ➢ shape restriction ➢ not suitable for high volume and routine quality control ➢ no real discretization in size ➢ no dimensional measurements

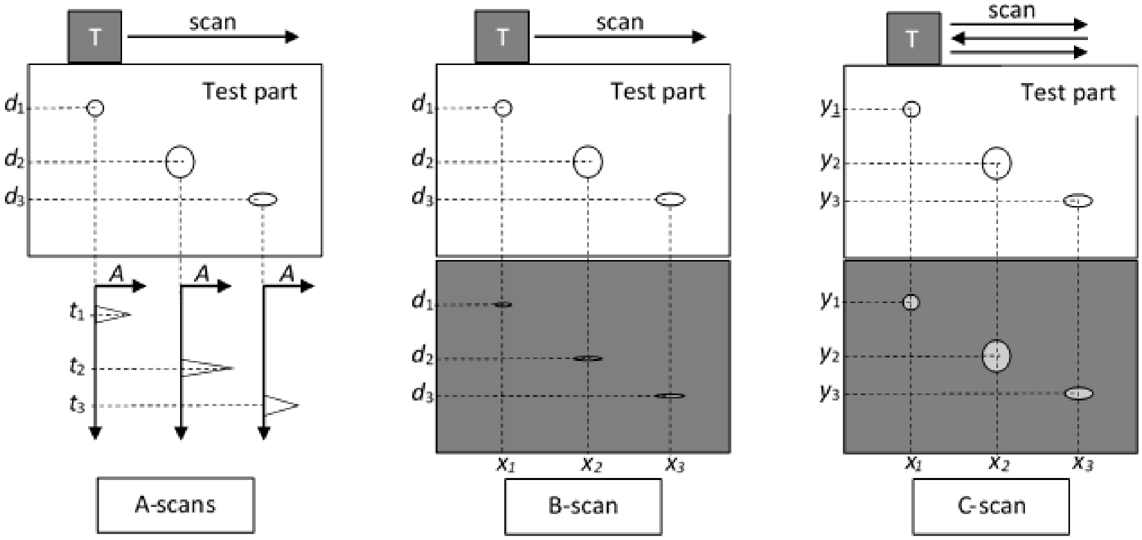


FIG. 10 Schematic representation of A, B, and C scans used in ultrasound inspection.

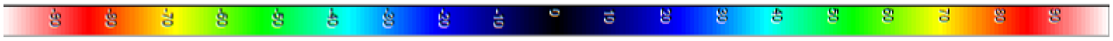


FIG. 11 Color code on the C-scan images (blue: low amplitudes, red: large amplitudes).

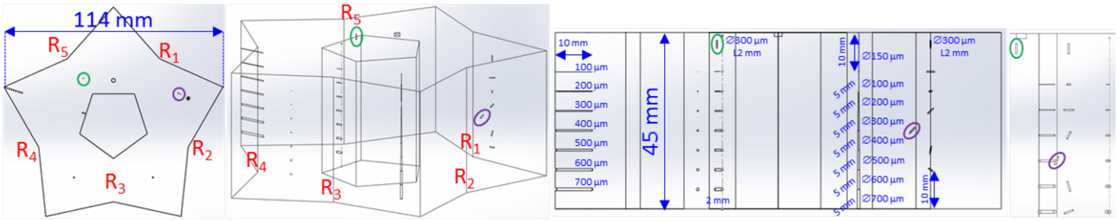


FIG. 12 CAD of the aluminum star artifact, design S1, seen from different sides

("R" refers to the Region in the star).

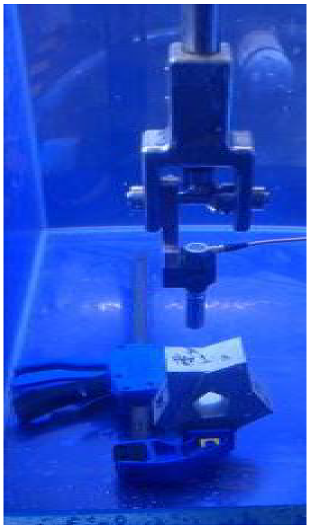


FIG. 13 CUT side inspection of the aluminum star artifact.

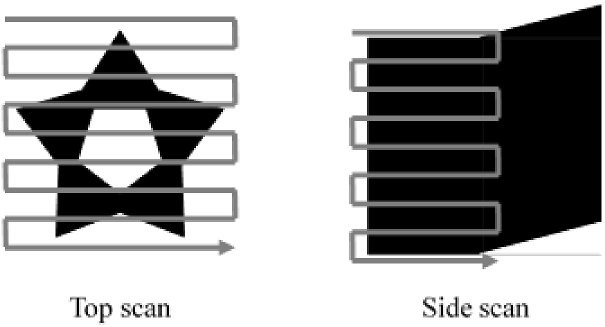


FIG. 14 Scanning strategies of the aluminum star artifact using the CUT method.

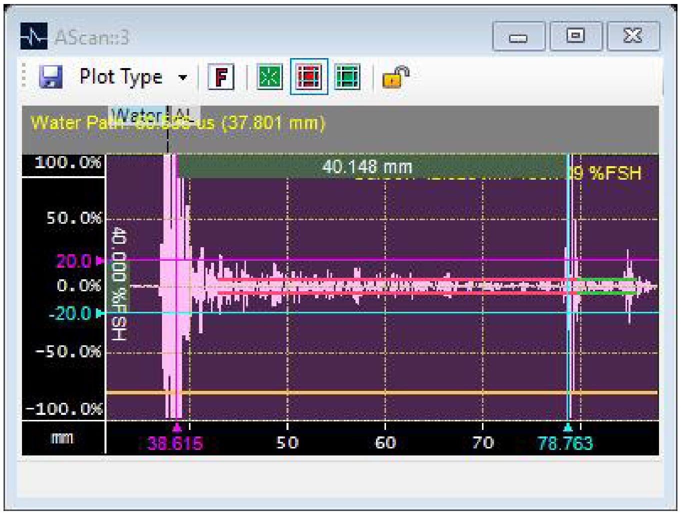


FIG. 15 Time-domain windows (red and green rectangles) defined on the A-scan plot of the reflected signal for data analysis of CUT.

TABLE 6 Transducer characteristics and experimental parameters used for the CUT inspection of the aluminum artifact.

Scan	Type	Frequency (MHz)	Focal legnth (mm)	Diameter (mm)	Wavelength (mm)	Spot size (mm)	Resolution (mm)
		γ	f	$\varnothing$	$\lambda=v_s/\gamma$	$1.22 \times f/\varnothing \times \lambda$	$0.707 \times 1.22 \times f/\varnothing \times \lambda$
from top	Ceramic	25	6	9.5	0.3	0.2	0.1
from side	Ceramic	10	75	9.5	0.6	6.1	4.3

Sound velocity in aluminum (mm/µs)	v_S	6.4

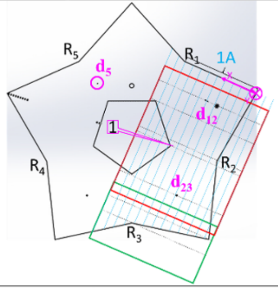
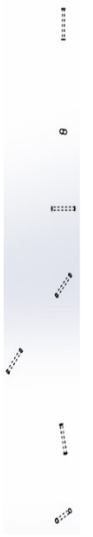
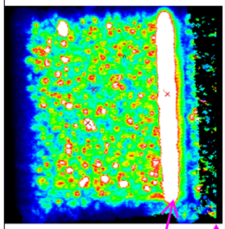
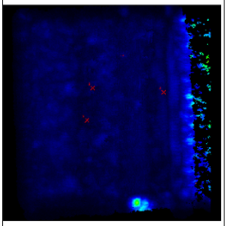
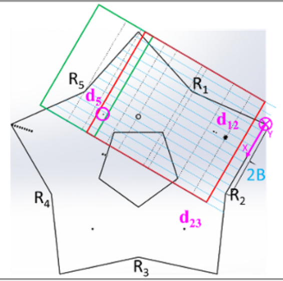
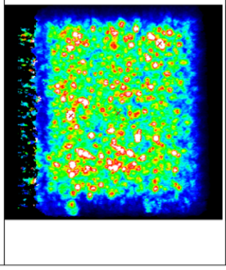
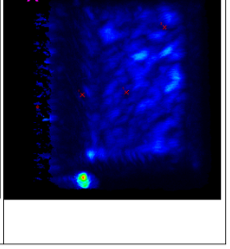
Scanned regions and time domain windows (red and green rectangles)	Defects in scanned regions	Scan side	C-scan near field (Data1: APA)	C-scan far field (Data2: APA)
 R_5 d_5 R_1 1A d_{12} 1 R_4 d_{23} R_2 R_3		1A		
 R_5 d_5 R_1 d_{12} 2B R_4 R_2 d_{23} R_3		2B		

FIG. 16 CUT inspection from sides 1A and 2B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 1A/2B, "data" refers to the time domain window, and "APA" stands for "Absolute Peak Amplitude".

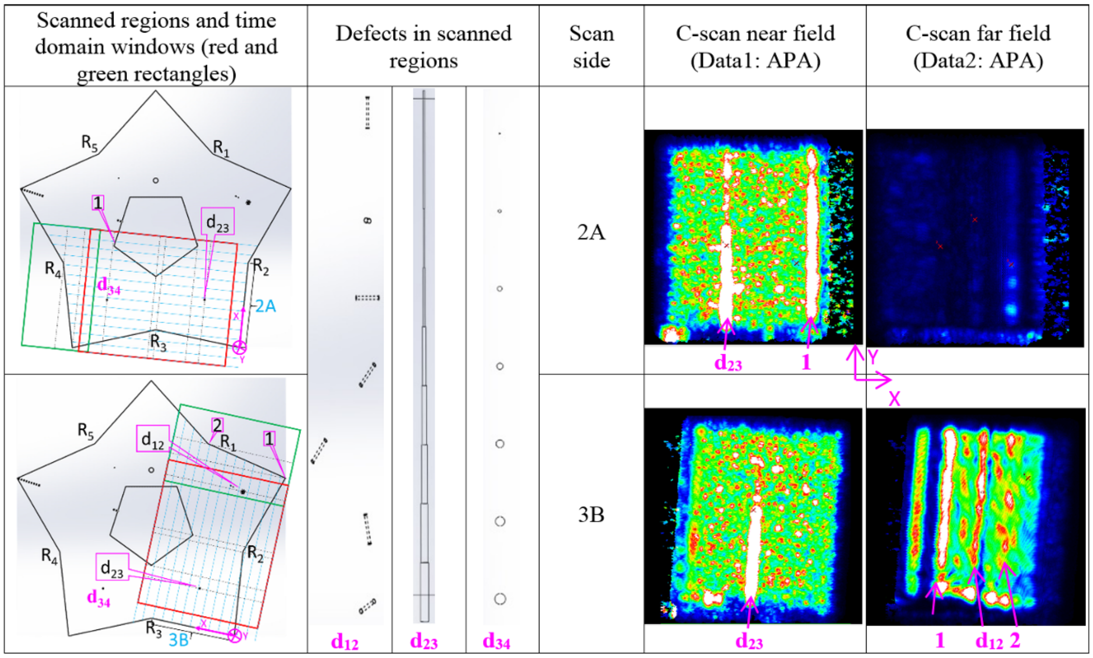

Scanned regions and time domain windows (red and green rectangles)
Defects in scanned regions
Scan side
C-scan near field (Data1: APA)
C-scan far field (Data2: APA)
R5
R1
1
d23
R4
d34
R2
2A
X
Y
R3
R5
d12
2
R1
1
R4
R2
d23
d34
R3
3B
X
Y
d12
d23
d34
2A
3B
d23
1
Y
X
d23
1
d12
2

FIG. 17 CUT inspection from sides 2A and 3B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 2A/3B, “data” refers to the time domain window, and “APA” stands for “Absolute Peak Amplitude”.

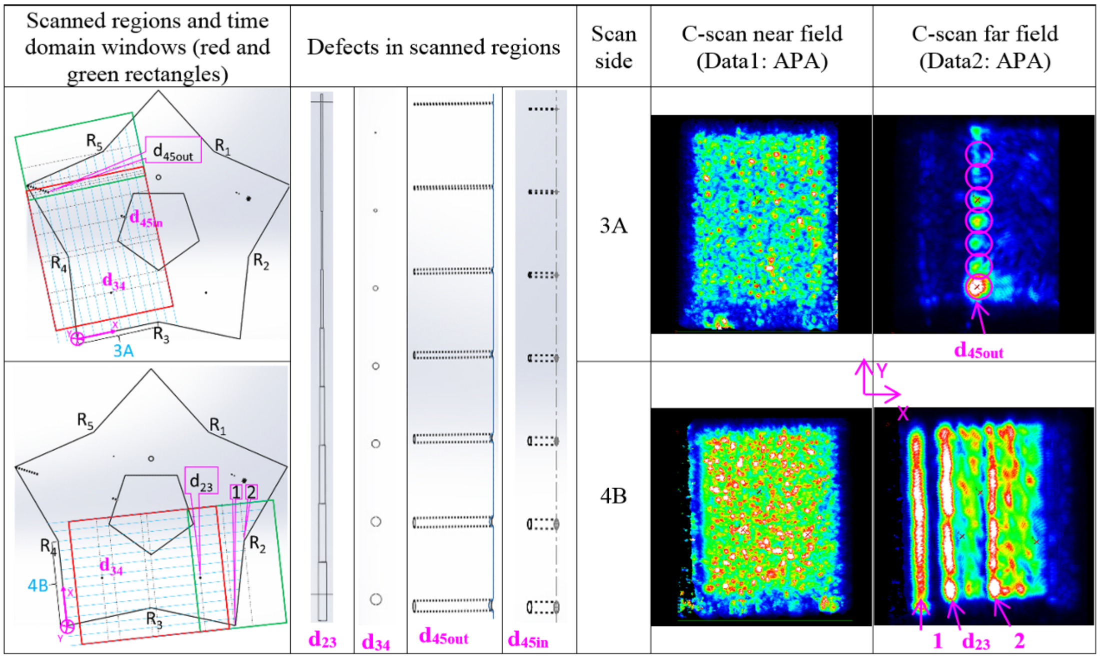

Scanned regions and time domain windows (red and green rectangles)
Defects in scanned regions
Scan side
C-scan near field (Data1: APA)
C-scan far field (Data2: APA)
R5
d45out
R1
d45in
R4
d34
R2
R3
3A
X
Y
d23
d34
d45out
d45in
3A
d45out
Y
X
R5
R1
d23
1
2
R4
d34
R2
4B
R3
4B
1
d23
2

FIG. 18 CUT inspection from sides 3A and 4B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 3A/4B, “data” refers to the time domain window, and “APA” stands for “Absolute Peak Amplitude”.

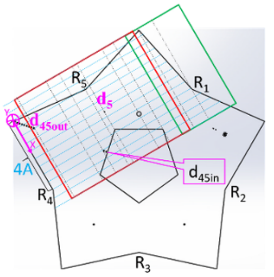
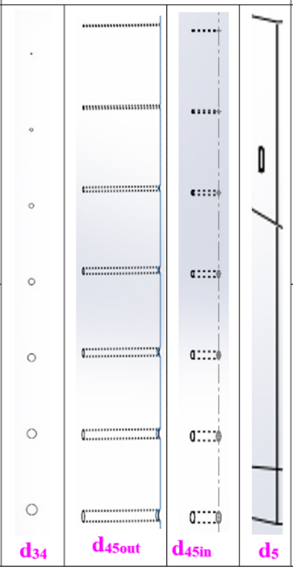
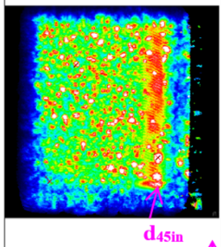
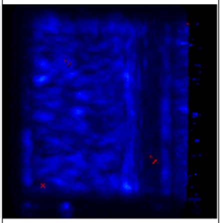
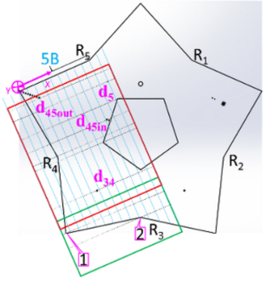
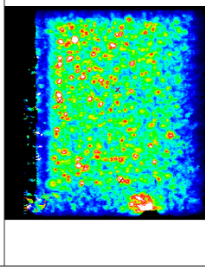
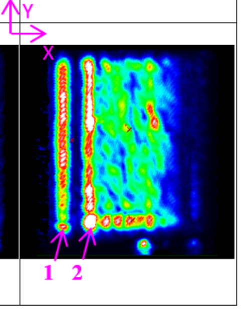
Scanned regions and time domain windows (red and green rectangles)	Defects in scanned regions	Scan side	C-scan near field (Data1: APA)	C-scan far field (Data2: APA)
 R R1 R5 d_5 d_{45out} d_{45in} 4A R4 R2 R3		4A		
 5B R5 R1 d_{45out} d_5 d_{45in} R4 R2 d_{34} 2 R3 1		5B		

FIG. 19 CUT inspection from sides 4A and 5B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 4A/5B, “data” refers to the time domain window, and “APA” stands for “Absolute Peak Amplitude”.

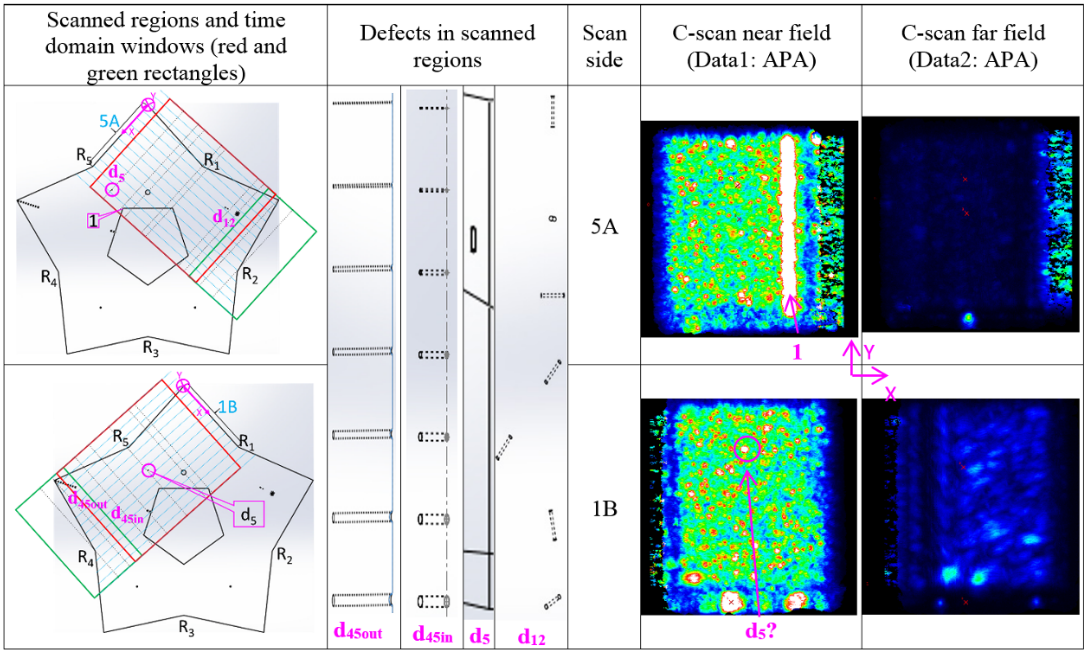

Scanned regions and time domain windows (red and green rectangles)
Defects in scanned regions
Scan side
C-scan near field (Data1: APA)
C-scan far field (Data2: APA)
5A
Y
X
R5
R1
d5
1
d12
R4
R2
R3
5A
1
Y
X
1B
R5
R1
d45out
d45in
d5
R4
R2
R3
1B
d5?
d45out
d45in
d5
d12

FIG. 20 CUT inspection from sides 5A and 1B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 5A/1B, “data” refers to the time domain window, and “APA” stands for “Absolute Peak Amplitude”.

Phased array ultrasonic testing (PAUT)

Principle of phased array ultrasonic testing (PAUT)

The principle of phased array ultrasonic testing (PAUT) is very similar to that of CUT. However, instead of consisting of a single piezoelectric transducer, the PAUT probe is composed of an array of piezoelectric transducers. Each transducer of this array can be pulsed independently with respect to the others to generate several waves that will interfere. The waves that are in phase will be added together (constructive interference), whereas the ones that are out of phase will cancel each other (destructive interference). This results in a unique wavefront constituting a synthesized focused beam that can be oriented (controllable angle), shaped (controllable focal distance and focal spot size) and swept electronically. Dynamic control of the beam enables examining the complex shapes across a range of different perspectives without moving the probe. The spatial resolution of the images is linked to the number of elements in the array. The larger the number of elements, the better the resolution.

It is possible to improve the performance of the above-mentioned standard PAUT system by implementing a data acquisition method called full matrix capture (FMC) with a post-processing reconstruction algorithm known as total focusing method (TFM) [11, 12, 13]. In the generic FMC acquisition method, a single transducer of the array is pulsed and the time domain signals (A-scans) are captured by all transducers in the array and stored for later processing. This is done for all transducers of the array. The number of data sets acquired is the square of the number of transducers. Thus, a 128-transducer probe will produce 16384 raw A-scans. The offline data processing uses the complete set of time-domain data from all combinations of transmitting and receiving transducers. The reconstructed image from the A-scans is equivalent to making coherent summations over all transducers in order to focus at each point in the target Region. Thus, FMC combined with TFM enables offline reconstruction of a more detailed image (with better spatial resolution, perspective, and defect definition) than a standard PAUT imaging method where all array elements create a unique wavefront to form a beam with a fixed focus.

The instrument [14] that was used to perform the PAUT of the star artifacts implements both FMC and TFM. However, it permits also the acquisition of standard PAUT data. The software associated with the hardware enables simultaneous A-B-C-scan displays, but also a D-scan which is similar to the C-scan but in a perpendicular direction. If the C-scan displays an image in the plane (XY), then the D-scan will display the image in the orthogonal plane (ZY). During the translation of the PAUT probe, the beam inside the probe is swept in the perpendicular direction to scan the entire surface. The color code on the images is the same as the one described above for the CUT images (Fig. 11).

Phased array ultrasonic testing (PAUT) – total focusing method (TFM) of the aluminum S1 and S2 star artifacts

The S1 and S2 designs of the aluminum star artifacts, with corresponding CAD shown in Figs. 12 and 21, could be both tested using PAUT-TFM. Only one linear probe acting as a transmitter/receiver was used to induce the longitudinal waves. The probe is 32 mm long and composed of 128 transducers separated from each other by 0.25 mm (center to center). The star artifact, immersed in a tank of water for the coupling (Fig. 22), was scanned only from one side of its branches (Fig. 22), from the top to the bottom of the star (Y axis – i.e., sideways as shown in Fig. 22). The transducers were operated at a center frequency of 10 MHz (bandwidth of 6 MHz), which corresponds to a wavelength of 0.62 mm in aluminum. The water path distance between the probe and the scanned surface of the artifact during the scan was 12.5 mm. Finally, the analog gain was adjusted to 40 dB to acquire the data. The digital gain can be subsequently adjusted on the reconstructed image (300 pixels x 400 pixels) offline.

The A, B, C, and D-scan presented in Fig. 23 through Fig. 26 for the S1 design, and in Fig. 27 through Fig. 31 for the S2 design, show interesting results:

- Three of the cylinders with different orientations can be seen from side 1A (Fig. 23). The orientations of the other cylinders prevent detection.
- Five vertical cylinders with different diameters can be detected from side 2A (Fig. 24) using the three optimized digital gains chosen to display the images. The cylinders with smaller diameters, down to 100 µm, can be detected with increased

gain (Fig. 24 bottom images) at the expense of the reduced quality of the image (lower signal to noise ratio).

- Three of the spheres of different diameters can be seen from side 3A (Fig. 25).
- Three of the horizontal cylinders of different diameters (10 mm long) located on an outside edge of the star can be seen from side 4A (Fig. 26). Their position on the edge hides the others.
- Some horizontal cylinders of different diameters located on an inside edge of the star are seen from side 4A (Fig. 26). They are located on an edge and are short (2 mm long), so they are difficult to differentiate from the reflection of the edge.
- It does not seem that we can see the single cylinder manufactured by mistake.

The defects are more visible on PAUT-TFM images from the S2 design than from the S1 design.

- Six cylinders with different orientations (instead of three in the S1 design) can be seen in Region 5 from side 4A (Fig. 31). This means that we can visualize defects down to 200 µm in diameter in various orientations.
- All vertical cylinders can be seen in Region 4 from the side labeled 3A (Fig. 30) like in the S1 design, though with increased gain. This means that we can visualize defects down to 100 µm in diameter.
- All spheres (instead of three in the S1 design) can be seen in Region 4 from the side labeled 3A (Fig. 29). This means that we can visualize defects down to 100 µm in diameter.
- The horizontal cylinders are easier to see when located inside the star artifact (Fig. 28), than when located outside the star artifact (Fig. 27). Four of the them are seen

in Fig. 28.

Thus, due to its high spatial resolution when used with FMC/TFM, the PAUT method reveals individual defects and even allows rough dimensional measurements. The interpretation of the images is easier than in CUT but is still not straight forward. However, like in CUT, the surface roughness did not prevent the part from being tested, and the size of the part was not an obstacle. Apart from the better resolution than in CUT, another real advantage of the PAUT-TFM methods is to be able to perform static inspection by steering the beam instead of scanning the probe. This is of particular interest for complex shapes that cannot be scanned.

The main benefits and drawbacks of the phased array ultrasonic testing (PAUT) – total focusing method (TFM) method are summarized below.

Benefits of the method	➢ possibility of inspection without scanning by steering the beam ➢ less restriction in shape than CUT ➢ less restriction in size than X-ray computed tomography ➢ fast scanning (real-time), faster than CUT ➢ few roughness restrictions ➢ able to identify the type and location of defects in the part ➢ discretization in size ➢ rough dimensional measurements
Drawbacks of the method	➢ part set-up required ➢ need experience ➢ requires more experience and training than CUT ➢ more expensive than CUT

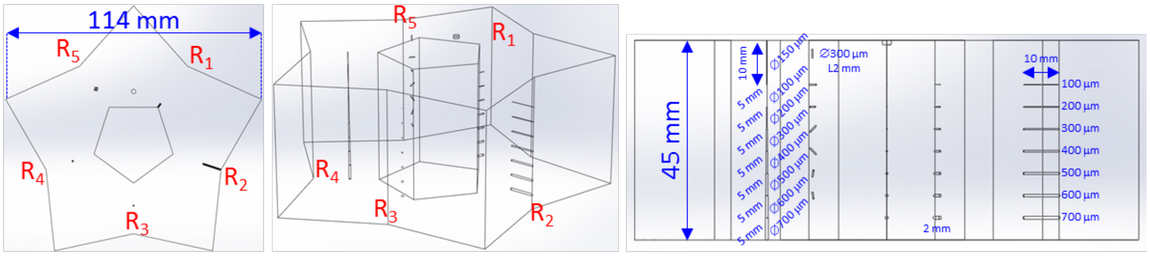


FIG. 21 CAD of the aluminum star artifact, design S2, seen from different sides

("R" refers to the Region in the star).

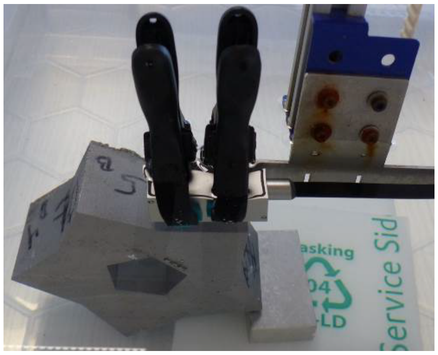


FIG. 22 PAUT-TFM side inspection of the aluminum star.

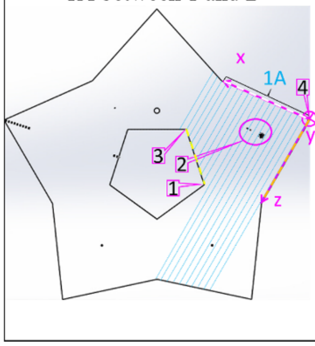
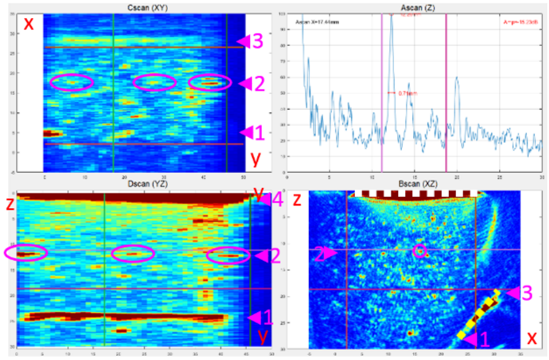
Scan side and region	Description of the defects		A, B, C, D-scans
1A between 1 and 2 	7 identical cylindrical cavities, with orientational offsets of 45 deg and 90 deg relative to first instance: 2 mm length 0.3 mm ∅, vertical 90^0 at the top (error in the CAD, misalignment of one of the cylinder)		

FIG. 23 PAUT-TFM inspection from side 5A of the S1 aluminum star artifact. The blue lines represent the longitudinal waves propagating from side 5A.

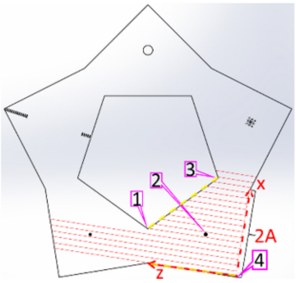
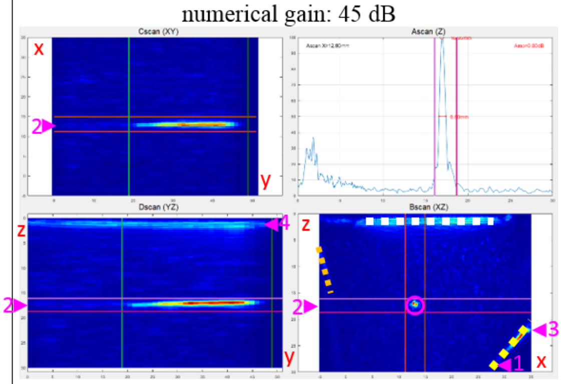
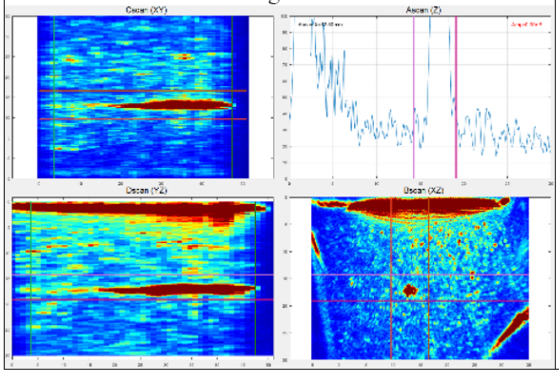
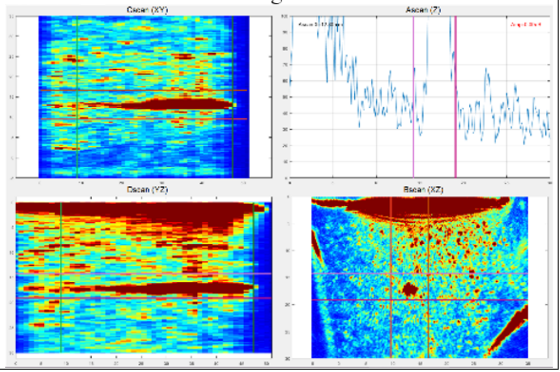
<table>
<tr><th>Scan side and region</th><th colspan="2">Description of the defects</th><th>A, B, C, D-scans</th></tr>
<tr><td>2A between 2 and 3
 </td><td>8 interconnected vertical cylinders, open at the outside at top and bottom of the star:
same height (5 mm except for top one 10 mm)
vary ∅ (150, 100, 200, 300 up to 700 µm from top to bottom)</td><td> </td><td>numerical gain: 45 dB
 </td></tr>
<tr><td colspan="3">numerical gain: 47 dB
 </td><td>numerical gain: 49 dB
 </td></tr>
</table>

FIG. 24 PAUT-TFM inspection from side 2A of the S1 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 2A. The two bottom scans correspond to higher digital gains.

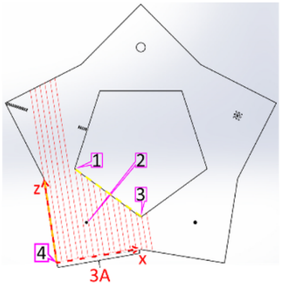
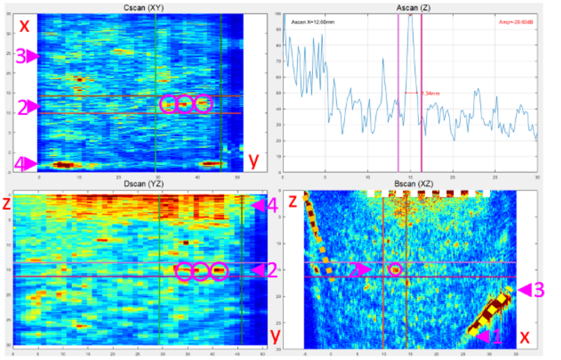
Scan side and region	Description of the defects		A, B, C, D-scans
3A between 3 and 4 	7 spherical cavities: 10 mm from the top 5 mm from the bottom 5 mm in between vary ∅ (100, 200 up to 700 µm from top to bottom)		

FIG. 25 PAUT-TFM inspection from side 3A of the S1 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 3A.

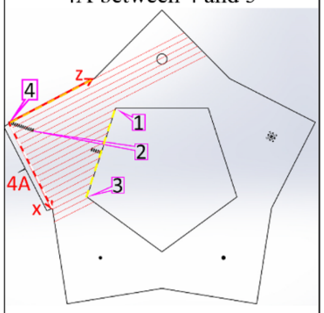
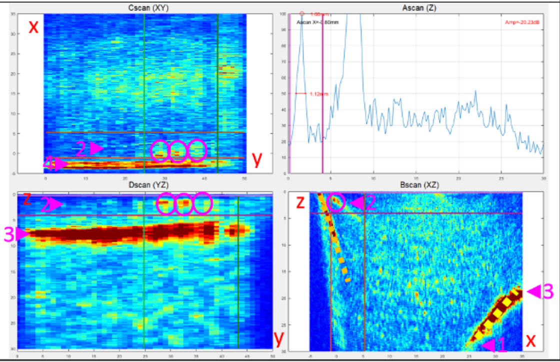
Scan side and region	Description of the defects		A, B, C, D-scans
4A between 4 and 5 	7 horizontal cylinders, open at the outside edge, on the outside of the star: 10 mm from the top 5 mm from the bottom 5 mm in between 2 mm height vary ∅ (100, 200 up to 700 µm from top to bottom)		

FIG. 26 PAUT-TFM inspection from side 4A of the S1 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 4A.

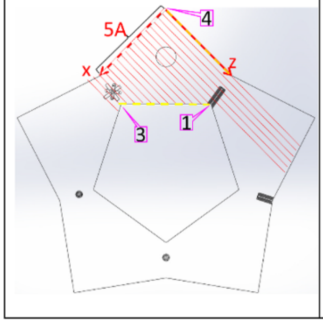
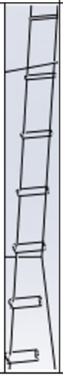
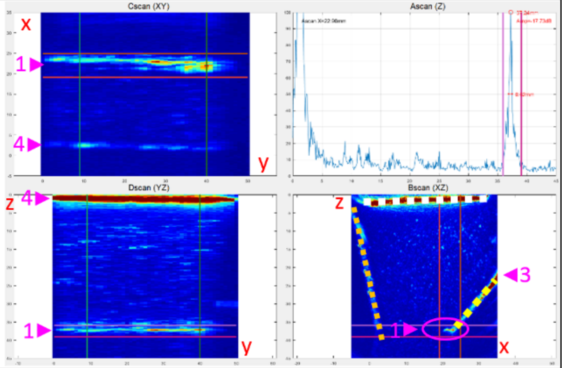
Scan side and region	Description of the defects	A, B, C, D-scans
5A region 1  	7 horizontal cylinders, open at the outside edge, on the inside of the star: 10 mm from the top 5 mm from the bottom 5 mm in between 2 mm height vary ∅ (100, 200 up to 700 µm from top to bottom) 	

FIG. 27 PAUT-TFM inspection from side 5A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 5A.

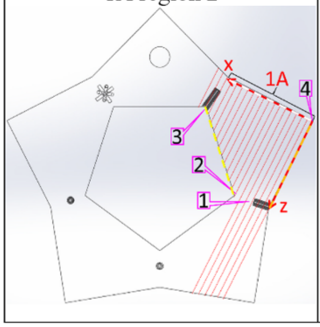
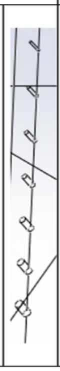
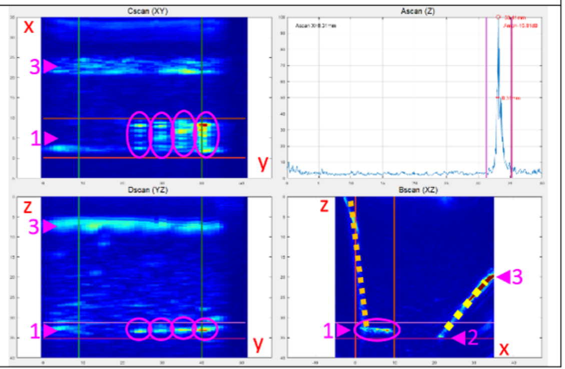
Scan side and region	Description of the defects	A, B, C, D-scans
1A region 2  	7 horizontal cylinders, open at the outside edge, on the outside of the star: 10 mm from the top 5 mm from the bottom 5 mm in between 2 mm height vary ∅ (100, 200 up to 700 µm from top to bottom) 	

FIG. 28 PAUT-TFM inspection from side 1A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 1A.

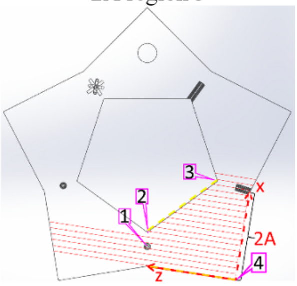
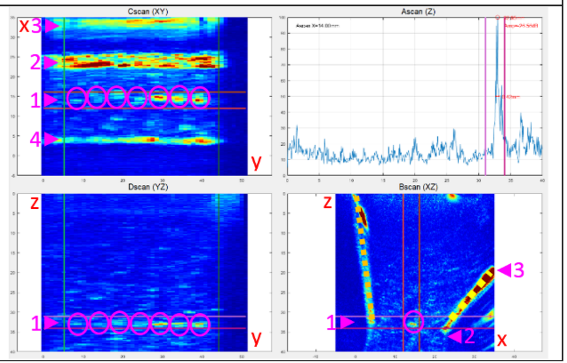
Scan side and region	Description of the defects	A, B, C, D-scans
2A region 3  	7 spherical cavities: 10 mm from the top 5 mm from the bottom 5 mm in between vary ∅ (100, 200 up to 700 µm from top to bottom) 	

FIG. 29 PAUT-TFM inspection from side 2A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 2A.

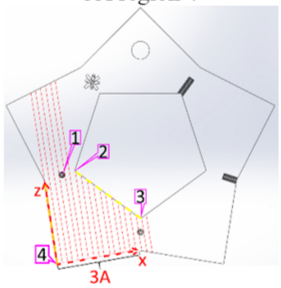
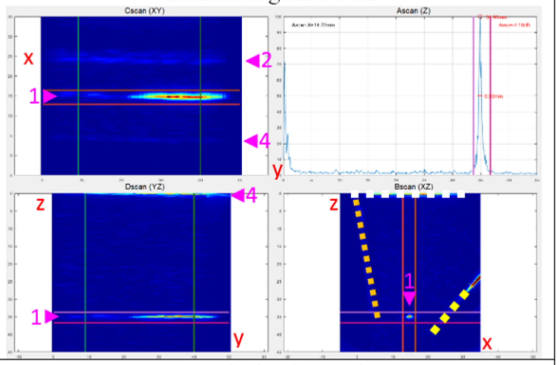
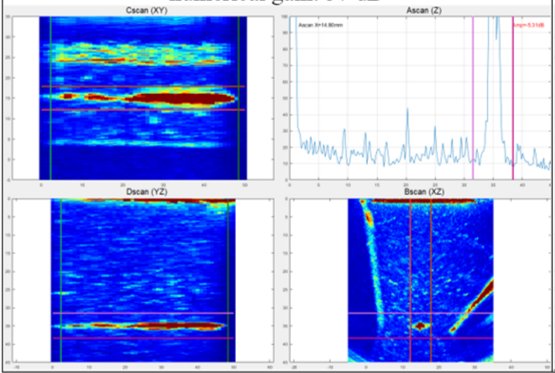
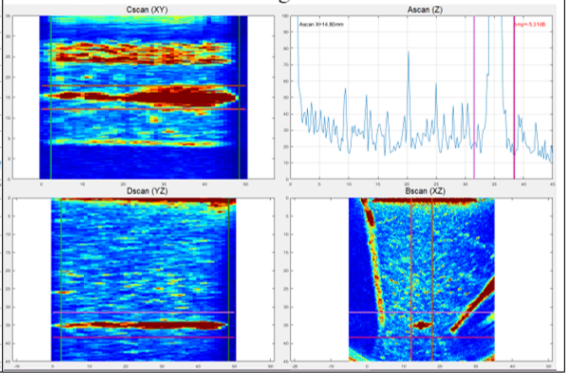
<table>
<tr><th>Scan side and region</th><th colspan="2">Description of the defects</th><th>A, B, C, D-scans</th></tr>
<tr><td>3A region 4

 </td><td>8 interconnected vertical cylinders, open at the outside at top and bottom of the star: same height (5 mm except for top one 10 mm) vary ∅ (150, 100, 200, 300 up to 700 µm from top to bottom)</td><td> </td><td>numerical gain: 50 dB

 </td></tr>
<tr><td colspan="3">numerical gain: 57 dB

 </td><td>numerical gain: 62 dB

 </td></tr>
</table>

FIG. 30 PAUT-TFM inspection from side 3A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 3A. The two bottom scans correspond to higher digital gains.

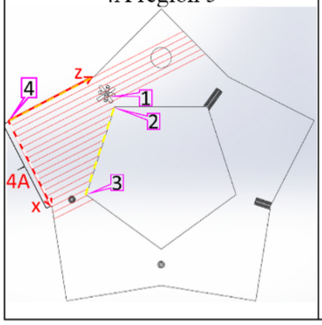
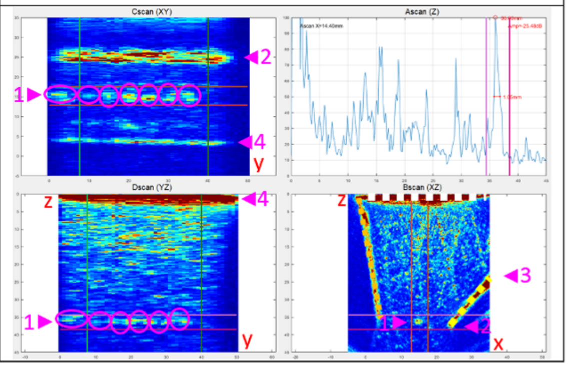
Scan side and region	Description of the defects	A, B, C, D-scans
4A region 5 	7 identical cylindrical cavities, with orientational offsets of 45 deg and 90 deg relative to first instance: 2 mm length 0.3 mm ∅, vertical 90⁰ at the top	

FIG. 31 PAUT-TFM inspection from side 5A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 5A.

X-ray computed tomography with synchrotron radiation (SRXCT)

Principle of X-ray computed tomography (XCT)

Tomography is a volumetric imaging method. It consists of probing a test part with electromagnetic or sound waves at different angular positions of the part, recording the interactions of these waves with the test part for each angle independently, and post-processing these data with a reconstruction algorithm to give a volumetric image. A tomographic device includes a wave source, a sensor that detects the effects of interaction between the waves and the test part (e.g., attenuation, reflection, diffusion, diffraction), and a translation stage with a rotating platform on which the test part is positioned. The most commonly used waves are X-rays, as the short wavelength allows the spatial resolution of the image to be better. The best-known method of X-Ray tomography is designated by the acronym XCT for "X-Ray computed tomography" [15, 16, 17]. In a laboratory XCT system, the X-Ray source consists of a high-voltage power supply (usually the voltage (*U*) ranges between 10 kV and 450 kV) and an X-Ray tube made of two electrodes: a filamentary cathode, usually of tungsten, and an anode, which is the target. By applying an electric current (I) to the cathode, electrons are extracted; then by applying a high potential difference (U) between the anode and the cathode, the electrons are accelerated to collide with the target, focused on a spot on the target, to emit X-Rays.

ASTM E1695-95 [18] describes two factors that affect the quality of a XCT image: geometrical unsharpness, which limits the spatial resolution, and random noise, which limits the contrast sensitivity. Spatial resolution influences the ability to detect small geometrical features in the object, while contrast sensitivity influences the ability to detect inhomogeneities in the object. The geometrical unsharpness is linked to the size of the focused spot. This size depends on the required power for the measurement ($P=UI$), which depends on the thickness and density of the test part. The thicker and higher density of the part, the higher the power required, which creates a larger focus spot. An increase in the size of the focus spot induces more geometrical unsharpness. Thus, a larger focus spot results in a decrease in the spatial resolution. The random noise is linked to the brilliance of the source, which is proportional to the number of photons produced per second. The higher the brilliance, the lower the random noise. Thus, a higher brilliance results in a higher contrast sensitivity.

The produced X-Ray beam is polychromatic (multi-wavelength) which is not ideal since artifacts (an artificial characteristic on the image) can appear on the image due to beam hardening (i.e., loss of the low energies in the source spectrum due to their absorption by the test part).

And finally, the beam is commonly conical. This shape causes the magnification (*mag*) of the part on the detector to depend on its position between the source and the detector. The closer the part to the source, the larger its magnification on the detector, which is commonly a flat panel composed of a matrix of pixels. This magnification and the size of a pixel of the detector ($pixel_{size}$)

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

will defined the spatial resolution of the image. Indeed, the spatial resolution is related to the size of the voxel (equivalent in three-dimensional (3D) of a pixel on the image) defined by:

$$voxel_{size} = pixel_{size}/mag \qquad (1)$$

The detector records the attenuated intensity of a ray, which is proportional to the thickness and density of material crossed. This enables acquiring two-dimensional (2D) images (i.e., greyscale images that show the radiation attenuation through the material) called projections. These projections, acquired over 360˚ around the test part, enable the reconstruction of the image in 3D.

Even though laboratory XCT is the more widespread X-Ray tomographic system, tomography is also performed with a synchrotron radiation X-ray source (SRXCT).

Principle of X-ray computed tomography with synchrotron radiation (SRXCT)

Synchrotron radiation is generated by electrons confined in a large loop and accelerated in a magnetic field. Due to the bending of their trajectory, the electrons accelerate or decelerate losing energy and thus produce electromagnetic radiation with a wide polychromatic spectrum from far infra-red to hard X-Rays. The use of a monochromator enables selecting a monochromatic X-Ray beam used to perform CT. The detector is generally a scintillator combined with a camera. Otherwise the SRXCT principle of image acquisition and reconstruction is the same as for laboratory XCT.

Compared to laboratory XCT, in addition to being monochromatic instead of polychromatic, the synchrotron beam is collimated instead of conical, and its brilliance is much higher than laboratory XCT. These are the three main differences between sources from laboratory XCT and synchrotron radiation. These differences are of paramount importance as they enable better spatial resolution, better contrast sensitivity, and fewer artifacts. Indeed, the benefit of a **monochromatic beam** is that it prevents artifacts on the image. The first benefit of a **parallel beam** is that the part does not need to be scanned over 360˚, but simply over 180˚.The information acquired is identical from one side to the other. The second benefit is that there is no focused spot which would create geometric unsharpness and thus decrease spatial resolution. The last benefit is that there is no magnification which means that the spatial resolution is defined only by the pixel size of the detector. The benefit of **high brilliance** is that it decreases random noise and thus increases contrast sensitivity.

X-ray computed tomography with synchrotron radiation (SRXCT) of the CoCr S2 star artifact

An XCT beamline (ID19) [19], from the European Synchrotron Radiation Facility (ESRF) located in Grenoble, France, was used to characterize the S2 design of the CoCr star artifact, with corresponding CAD shown in Fig. 32. The acquisition parameters are given in table 7 and the different images, processed with the ImageJ software, are displayed in Fig. 33 through Fig. 37.

The spatial resolution of the images is indisputably better than the ones from the UT methods. All defects are clearly located and imaged. Inspection of the 3D images also indicated an organic defect (hot tear or crack) at the base of the star (Fig. 38). The resolution of the images enables dimensional comparison with the CAD model (Fig. 39).

The main benefits and drawbacks of X-ray computed tomography (XCT) are summarized below.

Benefits of the method	➢ no restriction in shape ➢ no roughness restrictions ➢ able to identify the type and location of defects in the part ➢ discretization in size ➢ dimensional measurements
Drawbacks of the method	➢ part set-up required ➢ need experience ➢ not suitable for high volume and routine quality control ➢ more expensive than any other NDT methods

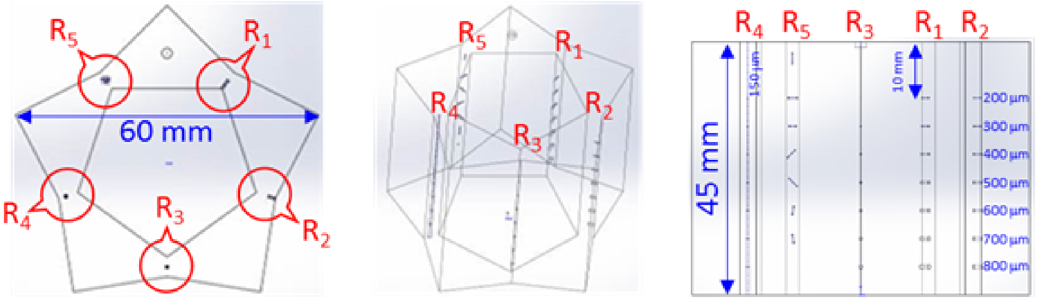


FIG. 32 CAD of the Co-Cr full-star artifact, S2 design, seen from different sides ("R" refer to the Region in the star).

TABLE 7 Experimental parameters used for the XCT inspection using synchrotron of the CoCr star.

Acquisition parameters		
Synchrotron radiation source	Energy (keV)	195
Detector	camera	PCO Edge 4.2
	scintillator	LuAg 2000
	image size (pixel)	2048x400
	pixel size (µm)	23.3
Acquisition mode (pixel)	Half acquisition 500	
Exposure time (ms)	15	
Image per projection	20	
Number of projections	5000	
Number of scans at different vertical locations	10	
Distance between sample and detector	13	

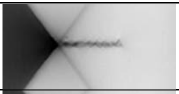
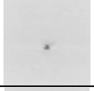
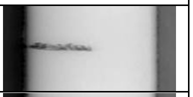
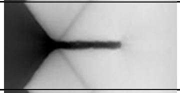
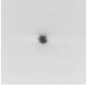
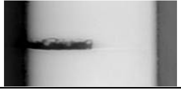
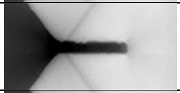
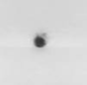
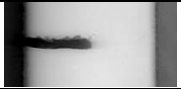
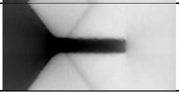
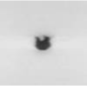
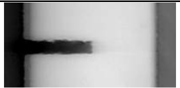
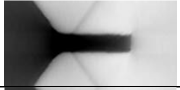
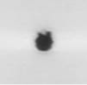
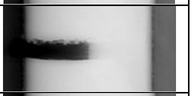
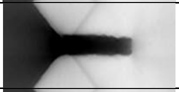
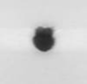
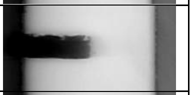
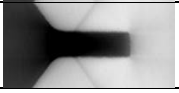
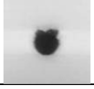
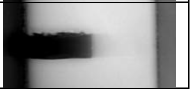
Region	CAD description	Defect characteristics		Top view	Face view	Side view
1	7 horizontal cylinders, open at the outside edge, on the inside of the star: 10 mm from the top 5 mm from the bottom 5 mm in between 2 mm height vary ∅, the smaller at the top	∅, µm	200			
			300			
			400			
			500			
			600			
			700			
			800			

FIG. 33 Star artifact images obtained by SRXCT for Region 1.

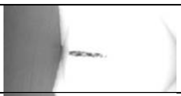
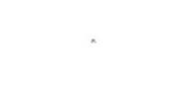
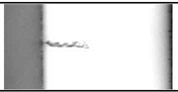
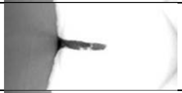
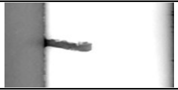
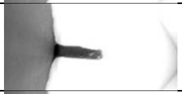
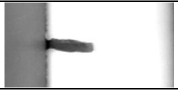
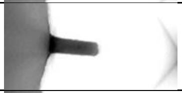
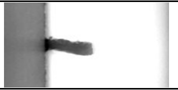
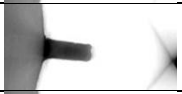
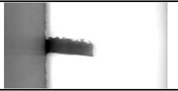
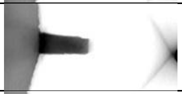
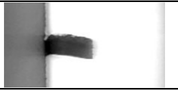
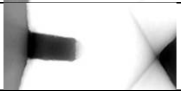
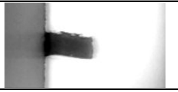
Region	CAD description	Defect characteristics		Top view	Face view	Side view
2	7 horizontal cylinders, open at the outside edge, on the outside of the star: 10 mm from the top 5 mm from the bottom 5 mm in between 2 mm height vary ∅, the smaller at the top	∅, µm	200			
			300			
			400			
			500			
			600			
			700			
			800			

FIG. 34 Star artifact images obtained by SRXCT for Region 2.

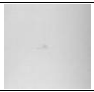
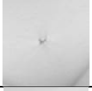
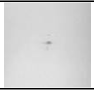
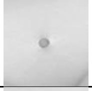
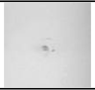
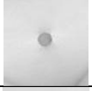
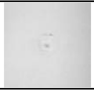
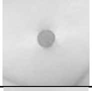
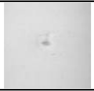
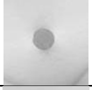
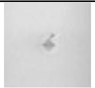
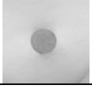
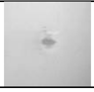
Region	Description	Defect characteristics		Top view	Face view
3	7 spherical cavities: 10 mm from the top 5 mm from the bottom 5 mm in between vary ∅, the smaller at the top	∅, µm	200		
			300		
			400		
			500		
			600		
			700		
			800		

FIG. 35 Star artifact images obtained by SRXCT for Region 3.

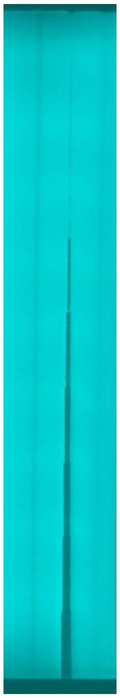
Region	Description	Defect characteristics		Top view	Face view
4	8 interconnected vertical cylinders, open at the outside at top and bottom of the star: vary heights vary ∅, the smaller at the top	∅, µm, and height, mm	∅=150 µm 10 mm height		
			∅=200 µm 5 mm height		
			∅=300 µm 5 mm height		
			∅=400 µm 5 mm height		
			∅=500 µm 5 mm height		
			∅=600 µm 5 mm height		
			∅=700 µm 5 mm height		
			∅=800 µm 5 mm height		

FIG. 36 Star artifact images obtained by SRXCT for Region 4.

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

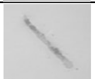
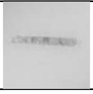
Region	Description	Defect characteristics		Top view	Face view
5	7 identical cylindrical cavities, with orientational offsets of 45 deg and 90 deg relative to first instance: 2 mm length 0.3 mm ∅, vertical 90^{0} at the top	Angle, deg, and distance from the top, mm	vertical 90^{0} 3 mm		
			horizontal 0^{0} (radial) 10 mm		
			horizontal 0^{0} (tangential) 15 mm		
			45^{0} 20 mm		
			-45^{0} 25 mm		
			45^{0} 30 mm		
			-45^{0} 35 mm		

FIG. 37 Star artifact images obtained by SRXCT for Region 5.

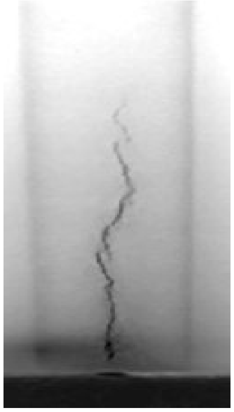


FIG. 38 Organic defect (hot tear or crack) in Region 2.

Obaton, Anne Francoise; Butsch, Bryan; McDonough, Stephen; Laroche, Nans; Gaillard, Yves; Tarr, Jared; Bouvet, Patrick; Cruz, Rodolfo; Donmez, Alkan. "Evaluation of Nondestructive Volumetric Testing Methods for Additively Manufactured Parts." Paper presented at ASTM Symposium on Structural Integrity of Additive Manufactured Parts, Washington, DC, US. November 06, 2018 - November 08, 2018.

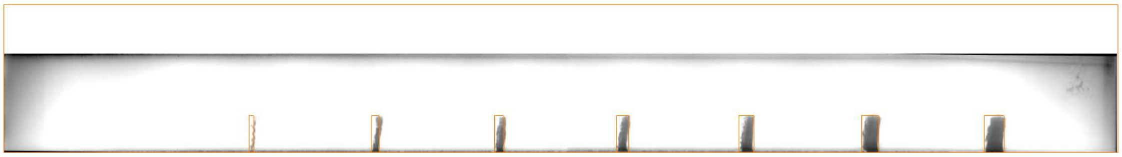


FIG. 39 Superposition of the CAD and the X-Ray scan in Region 2.

CONCLUSION

We have investigated four volumetric NDT methods as potential alternatives to XCT for faster and cheaper characterization for routine control and inspection of additively manufactured metal parts: two whole-body inspection methods using acoustic waves (RAM and PCRT) and two selective inspection methods using ultrasound waves (CUT and PAUT-TFM). The two whole-body inspection methods gave similar results. The parts with defects could be separated from the parts without defects, but the parts with defects could not be sorted based on the number of defects they contained. The results of the two selective inspection methods were significantly different. The CUT method allowed detection of most of the defects with low spatial resolution, though defects 5 mm apart could not be discriminated. Furthermore, the defects were not detected from every direction; successful results depended on the side inspected . The PAUT-TFM resulted in higher resolution images than CUT, allowing detection and separation of nearly all defects for a given side inspection. As a comparison we performed SRXCT which gives better spatial resolution, better contrast sensitivity, and fewer artifacts than laboratory XCT.

The whole-body inspection methods allowed only the identification of defective parts, whereas the selective inspection methods allowed identification of the type of the defect. In applications like routine inspections, it is sufficient to determine that a part is non-conforming in order to reject it. Then if localization and size of defects are needed to evaluate the defect criticality and impact on part performance, a secondary evaluation could use a selective inspection method.

Further volumetric NDT methods, such as non-linear acoustic methods, should be investigated in order to boost the application of additive manufacturing in critical sectors and to guarantee the reliability of parts for critical applications.

ACKNOWLEDGMENTS

We would like to particularly thank Wilson Vesga, Ben Dutton (convenor of the JG59), and David Ross-Pinnock from MTC for their investment in the group and the valuable information provided to help us. We are also grateful to Zodiac-Aerospace for building the star artifacts in aluminum, to Sofiane Terzi and Barbara Fayard from Novitom for the SRXCT inspection, and to Julieanne Heffernan from Vibrant for the PCRT inspection.

Disclaimer

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] Obaton, A-F., Lê, M-Q., Prezza, V., Marlot, D., Delvart, P., Huskic, A., Senck, S., Mahé, E., Cayron, C., “Investigation of new volumetric non-destructive techniques to characterise additive manufacturing parts,” *Welding in the World*, Vol. 62, Issue 5, 2018, pp. 1049-1057,

https://doi.org/10.1007/s40194-018-0593-7

[2] Perraud, J.B., Obaton, A.F., Bou-Sleiman, J., Recur, B., Balacey, H., Darrack, F., Guillet, J.P., Mounaix, P., "THz imaging and tomography as efficient instruments for testing polymer additive manufacturing objects," *Applied Optics,* Vol. 55, Issue 13, 2016, pp. 3462-3467, https://doi.org/10.1364/AO.55.003462

[3] Todorov, E., Spencer, R., Gleeson, S., Jamshidinia, M., and Shawn M. Kelly, "America Makes: National Additive Manufacturing Innovation Institute (NAMII) Project 1: Nondestructive Evaluation (NDE) of Complex Metallic Additive Manufactured (AM) Structures," Interim Report, June 2014.

[4] ASTM E2001 − 13: Standard Guide for Resonant Ultrasound Spectroscopy for Defect Detection in Both Metallic and Non-metallic Parts, ASTM International, West Conshohocken, PA, 2013, www.astm.org,

http://dx.doi.org/10.1520/A0252-10

[5] ASTM E2534 – 15: Standard Practice for Process Compensated Resonance Testing Via Swept Sine Input for Metallic and Non-Metallic Parts, ASTM International, West Conshohocken, PA, 2015, www.astm.org,

http://dx.doi.org/10.1520/A0252-10

[6] Stultz, G, Bono, R, Schiefer, M, "Fundamentals of Resonant Acoustic Method NDT", http://www.modalshop.com/techlibrary/Fundamentals%20of%20Resonant%20Acoustic%20Method%20NDT.pdf (accessed Sept. 26, 2018)..

[7] NIST, National Institute of Standards and Technology, "Detection of Outliers," https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm (accessed Sept. 26, 2018).

[8] Woodall, W. H., Koudelik, R., Tsui, K-L., Bum Kim, S., Stoumbos, Z. G., Carvounis C. P. A,

"Review and Analysis of the Mahalanobis—Taguchi System, Technometrics," Vol.45, No.1, 2003, pp 1-15,

https://doi.org/10.1198/004017002188618626

[9] Biedermann, E., Heffernan, J., Mayes, A., Gatewood, G., Jauriqui, L., Goodlet, B., Pollock, T., Torbet, C., Aldrin, J. C., Mazdiyasni, S., "Process compensated resonance testing modeling for damage evolution and uncertainty quantification,", *AIP Conference Proceedings* 1806, 090005, 2017,

https://doi.org/10.1063/1.4974649

[10] ASTM E1001 – 16: Standard Practice for Detection and Evaluation of discontinuities by the Immersed Pulse-Echo Ultrasonic Method Using Longitudinal Waves, ASTM International, West Conshohocken, PA, 2016, www.astm.org,

http://dx.doi.org/10.1520/A0252-10

[11] Karaman, M. Li, P.-C. and O'Donnell, M., "Synthetic aperture imaging for small scale systems," *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, Vol. 42, No.3, 1995, pp. 429–442,

http://dx.doi.org/10.1109/58.384453

[12] Holmes, C. Drinkwater, B. W. and Wilcox, P. D., "Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation," *NDT&E International*, Vol. 38, No. 8, 2005, pp.701–711,

http://dx.doi.org/10.1016/j.ndteint.2005.04.002

[13] Spencer, R., Sunderman, R., Todorov, E., "FMC/TFM experimental comparisons," *44th Annual Review of Progress in Quantitative Nondestructive Evaluation*, Vol. 37, 2018,

https://doi.org/10.1063/1.5031512

[14] Carcreff, E., Dao, G., and Braconnier. D., "Total focusing method for flaw characterization in homogeneous media," *Fourth International Symposium on Nondestructive Characterization of Materials*, Marina Del Rey, USA, June 22-26, 2015, https://www.researchgate.net/profile/Ewen_Carcreff/publication/282654039_Total_focusing_method_imaging_for_flaw_characterization_in_homogeneous_media/links/56162ac708ae4ce3cc65a588/Total-focusing-method-imaging-for-flaw-characterization-in-homogeneous-media.pdf

[15] Obaton, A-F., Fain, J., Djemaï, M., Meinel, D., Léonard, F., Mahé, E., Lécuelle, B., Fouchet, J-J., Bruno, G., "In vivo XCT bone characterization of lattice structured implants fabricated by additive manufacturing: a case report," *Heliyon*, Vol. 3, 2017, http://dx.doi.org/10.1016/j.heliyon.2017.e00374

[16] Hermanek, P., Carmignato, S., "Reference object for evaluating the accuracy of porosity measurements by X-ray computed tomography," *Case Studies in Nondestructive Testing and Evaluation*, Vol. 6, 2016, pp.122–127, http://dx.doi.org/10.1016/j.csndt.2016.05.003

[17] Kim, F.H., Moylan, S.P., Garboczi, E.J., Slotwinski, J.A., "Investigation of pore structure in cobalt chrome additively manufactured parts using X-ray computed tomography and three-dimensional image analysis," *Addit. Manuf.*, Vol. 17, 2017, pp. 23–38, http://dx.doi.org/10.1016/j.addma.2017.06.011

[18] ASTM E1695 – 95: Standard Test Method for Measurement of Computed Tomography (CT) System Performance, ASTM International, West Conshohocken, PA, 2016, www.astm.org, http://dx.doi.org/10.1520/E1695-95R13

[19] ESRF, European Synchrotron Radiation Facility, "ID19 - Microtomography beamline," https://www.esrf.eu/home/UsersAndScience/Experiments/StructMaterials/ID19.html (accessed

Sept. 26, 2018).

List of Figure Captions

FIG. 1 Schematics of the AM star artifacts (design S1 and S2) proposed by JG59 (where R is a Region in the artifact, h and a define the height and width of the artifact, respectively, and numerical values are given in Table 2).

FIG. 2 Photos of the manufacturing of the Co-Cr (a and b) and SS (c and d) star artifacts, during the build (a and c) and after the build on the machine platform (b and d).

FIG. 3 Photos of the manufactured star artifacts, from left to right: Al, SS, and Co-Cr.

FIG. 4 The Modal Shop RAM set-up (a) and inspection of the half-size SS star artifacts using the set-up with the automatic hammer (b).

FIG. 5 Computer Aided Design (CAD) of the SS half-size star artifact, S2 design, seen from different sides ("R" refers to the Region in the star).

FIG. 6 Typical half-size SS star artifact RAM spectrum showing the subset of the six resonant peaks used for comparison criteria.

FIG. 7 PCRT set-up used for inspection of the half-size SS star artifacts.

FIG. 8 Typical half-size SS star artifact PCRT spectrum. The green plots represent four good parts and the red ones are four bad parts.

FIG. 9 PCRT overall Pass/Fail results for the 74 half-size SS star artifacts based on MTS and Bias scores.

FIG. 10 Schematic representation of A, B, and C scans used in ultrasound inspection.

FIG. 11 Color code on the C-scan images (blue: low amplitudes, red: large

amplitudes).

FIG. 12 CAD of the aluminum star artifact, design S1, seen from different sides (“R” refers to the Region in the star).

FIG. 13 CUT side inspection of the aluminum star artifact.

FIG. 14 Scanning strategies of the aluminum star artifact using the CUT method.

FIG. 15 Time-domain windows (red and green rectangles) defined on the A-scan plot of the reflected signal for data analysis of CUT.

FIG. 16 CUT inspection from sides 1A and 2B of the aluminum star artifact. The red/blue lines represent the longitudinal waves propagating from sides 1A/2B, “data” refers to the time domain window, and “APA” stands for Absolute Peak Amplitude”.

FIG. 16 CUT inspection from sides 1A and 2B of the aluminum star artifact. The red/blue lines represent the longitudinal waves propagating from sides 1A/2B, “data” refers to the time domain window, and “APA” stands for Absolute Peak Amplitude”.

FIG. 17 CUT inspection from sides 2A and 3B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 2A/3B, “data” refers to the time domain window, and “APA” stands for Absolute Peak Amplitude”.

FIG. 18 CUT inspection from sides 3A and 4B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 3A/4B, “data” refers to the time domain window, and “APA” stands for Absolute Peak Amplitude”.

FIG. 19 CUT inspection from sides 4A and 5B of the aluminum star artifact. The blue lines represent the longitudinal waves propagating from sides 4A/5B, “data” refers to the time domain window, and “APA” stands for Absolute Peak Amplitude”.

FIG. 20 CUT inspection from sides 5A and 1B of the aluminum star artifact. The

blue lines represent the longitudinal waves propagating from sides 5A/1B, "data" refers to the time domain window, and "APA" stands for Absolute Peak Amplitude".

FIG. 21 CAD of the aluminum star artifact, design S2, seen from different sides ("R" refers to the Region in the star).

FIG. 22 PAUT-TFM side inspection of the aluminum star.

FIG. 23 PAUT-TFM inspection from side 5A of the S1 aluminum star artifact. The blue lines represent the longitudinal waves propagating from side 5A.

FIG. 24 PAUT-TFM inspection from side 2A of the S1 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 2A. The two bottom scans correspond to higher digital gains.

FIG. 25 PAUT-TFM inspection from side 3A of the S1 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 3A.

FIG. 26 PAUT-TFM inspection from side 4A of the S1 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 4A.

FIG. 27 PAUT-TFM inspection from side 5A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 5A.

FIG. 28 PAUT-TFM inspection from side 1A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 1A.

FIG. 29 PAUT-TFM inspection from side 2A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 2A.

FIG. 30 PAUT-TFM inspection from side 3A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 3A. The two bottom scans correspond to higher digital gains.

FIG. 31 PAUT-TFM inspection from side 5A of the S2 aluminum star artifact. The red lines represent the longitudinal waves propagating from side 5A.

FIG. 32 CAD of the Co-Cr full-star artifact, S2 design, seen from different sides (“R” refer to the Region in the star).

FIG. 33 Star artifact images obtained by SRXCT for Region 1.

FIG. 34 Star artifact images obtained by SRXCT for Region 2.

FIG. 35 Star artifact images obtained by SRXCT for Region 3.

FIG. 36 Star artifact images obtained by SRXCT for Region 4.

FIG. 37 Star artifact images obtained by SRXCT for Region 5.

FIG. 38 Organic defect (hot tear or crack) in Region 2.

FIG. 39 Superposition of the CAD and the X-Ray scan in Region 2.

A Study of Timing Constraints and SAS Overload of SAS-CBSD Protocol in the CBRS Band

Anirudha Sahoo, Naceur El Ouni* and Vineet Shenoy*
National Institute of Standards and Technology
Email: anirudha.sahoo@nist.gov, naceuramine@gmail.com, vineet.r.shenoy@rutgers.edu

***Abstract*—In the Citizens Broadband Radio Service (CBRS) band, the Federal Communications Commission (FCC) has set stringent timing constraints for the lower tier users to vacate the channel on which an incumbent shipborne radar appears. The standards body formulating various specifications for the CBRS operation has taken these timing constraints into consideration in the Spectrum Access System (SAS) – CBRS Device (CBSD) protocol. A transmitting CBSD continually heartbeats with its SAS. When required, the SAS sends commands to vacate a channel through these heartbeat messages. In this paper, we study the impact of the heartbeat interval on the CBRS system in terms of meeting the FCC timing constraints. We also study how the heartbeat interval can overload a SAS and how it can be used to determine the number of CBSDs a SAS can serve without causing unnecessary suspension of CBSD transmissions. We show the tradeoff between using a short heartbeat interval to meet the timing constraint early and the number of CBSDs that can be served by a SAS without causing unnecessary suspension of CBSD transmissions.**

I. Introduction

The Federal Communications Commission (FCC) recently published rules for commercial operators to use the 3.5 GHz band, also termed as Citizens Broadband Radio Service (CBRS) band, on a priority based sharing [1]. A Spectrum Access System (SAS) manages the use of spectrum in the CBRS band. A CBRS device (CBSD) has to get authorization from its managing SAS to use the spectrum and must vacate the spectrum when instructed by its SAS to do so. The communication protocol between a SAS and a CBSD has been standardized by the Spectrum Sharing Committee (SSC) of the Wireless Innovation Forum (WInnForum), commonly referred to as the *SAS-CBSD* protocol [2]. The CBRS band operationally is a three tiered system. The incumbents operate in tier-1 with the highest priority. The commercial operators may operate in tier-2 with medium priority or in tier-3 with lowest priority. One of the incumbents in the CBRS band is the shipborne Navy radars. When an incumbent Navy radar appears within the harmful interference range of deployed CBSDs, a SAS has to carefully identify the CBSDs which should be instructed to vacate the channel (in which the radar is operating) such that the interference to the radar receiver falls below a given threshold. The presence of Navy radars is detected by a set of sensors, known as Environment Sensing Capability (ESC) sensors, usually deployed along the coast of the US. The FCC rules have stringent timing requirements for the SAS and the CBSDs to protect incumbent radars from harmful interference. Once a SAS is notified of the presence of a Navy radar by an ESC, the SAS must ensure that the CBSDs, which may cause harmful interference to the radar, have vacated the channel within 300 s [1]. Once a CBSD has been instructed by the SAS to vacate the channel, the CBSD must do so within 60 s [1].

The SAS-CBSD protocol specification has been carefully designed to ensure that the timing requirements set forth by the FCC rules can be met [2]. A CBSD requests authorization to transmit on a channel (frequency range) by sending a grant request to the SAS. Each active grant in the SAS-CBSD protocol has a heartbeat mechanism through which the CBSD knows that the SAS is alive and vice-versa. A CBSD has to send a Heartbeat Request message to its managing SAS periodically for each of its active grants. How frequently a CBSD should send a Heartbeat Request message is decided by the SAS by setting the *heartbeatInterval* parameter. An actively transmitting Grant has a timer called *transmitExpireTime* timer. If and when this timer expires, the CBSD will have 60 s to turn off its transmission. Hence, to meet the end to end timing requirement of 300 s, the value of transmitExpireTime timer should not be more than 240 s for actively transmitting grants. The transmitExpireTime timer thus guards against violation of timing constraints in case communication between SAS and CBSD fails. A SAS can also change various protocol parameter values through the heartbeat messages to control the system timings. For example, the transmitExpireTime timer value and heartbeatInterval value can be changed through the hearbeat messages. All the request messages originate from the CBSDs and the SAS reponds to them with the corresponding response messages. Thus, when a SAS gets the notification from an ESC that an incumbent has appeared on a channel, the SAS has to wait for the next Heartbeat Request message to instruct the CBSD to vacate the channel via the corresponding Heartbeat Response message. There is a tradeoff between heartbeatInterval and processing load on the SAS. When the heartbeatInterval is small, the SAS can instruct the CBSD to vacate a channel sooner, but the processing load on the SAS is higher since the heartbeat rate is higher. On the other hand, when the heartbeatInterval is large, the processing load on the SAS is lower, but the SAS has to wait longer, on the average, to intsruct a CBSD to vacate a channel. The latter configuration can push the time a SAS takes to ensure that a CBSD vacates the channel close to the time limit of 300 s.

A SAS also needs to have adequate provisioning of processing resources so that it can handle CBSD request messages in

*The author was at National Institute of Standards and Technology when this work was done.

Sahoo, Anirudha; El Ouni, Naceur; Shenoy, Vineet. "A Study of Timing Constraints and SAS Overload in the CBRS Band using SAS-CBSD Protocol." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Wikola, HI, US. December 09, 2019 - December 13, 2019.

a timely manner. Since Heartbeat Request messages are the most frequent messages a SAS receives, their load dominates in determining the processing resource requirement of a SAS. If a Heartbeat Request message is dropped or sufficiently delayed due to lack of processing resources at a SAS, then the transmitExpireTime timer at the CBSD would expire. This would force the CBSD to suspend its transmission. This obviously is an unnecessary timeout caused by poor provisioning of processing resources in the SAS which leads to inefficient use of the spectrum.

Various efforts towards spectrum sharing in the US are summarized in [3]. The article also presents an example SAS architecture that facilitates tiered services in the CBRS band. The FCC rules on the CBRS operation were published in [1]. Based on the FCC rules, the Spectrum Sharing Committee (SSC) of Wireless Innovation Forum (WInnForum) has developed requirements for commercial operation in the CBRS band [4]. The WInnForum SSC has also developed the specification for the SAS-CBSD protocol [2]. Besides these, there has not been any work on the SAS-CBSD protocol that is publicly available.

To the best of our knowledge, there is no study on the SAS-CBSD protocol that is available in the public domain. Hence, the motivation behind this paper is two fold. First, we want to study the impact of heartbeatInterval on the performance of the CBRS system in terms of meeting the end to end timing constraint. Second, we want to study how message overload on a SAS impacts the performance of the CBRS system in terms of unnecessary timeouts of the transmitExpireTime timer which leads to suspension of CBSD transmission. To achieve these goals we have developed a basic simulator of the SAS-CBSD protocol. The protocol is implemented using the Omnet++ discrete event simulator package [5]. Our experiments were run for a CBRS system having very high number of CBSDs (up to 50 000) in the system. So, we believe the scale of simulation is similar to what we may see in practice. Our results show that a moderate mean hearbeatInterval of 150 s provides a good balance between handling the processing load and meeting the timing constraint. For a given SAS message service rate, our study also provides approximately how many CBSDs the SAS can serve without causing unnecessary timeouts (which would cause CBSDs to stop their transmissions). We present an approximate method of calculating the number of CBSDs a SAS can manage for a given request service rate, without causing unnecessary transmitExpireTime timeouts. This model can be used, for example, by SAS providers to decide when a new SAS process should be spawned to serve increasing number of CBSDs.

II. Simulation of SAS-CBSD Protocol

A. *Brief Description of the SAS-CBSD Protocol*

The SAS-CBSD protocol has two state machines: *Registration State Machine* and *Grant State Machine*. A CBSD starts out in the Unregistered state. When it registers with its SAS, it goes through the Registration State Machine as shown in Figure 1. If the registration is successful it transitions into Registered state. Once the CBSD is in Registered state, it is allowed to ask for spectrum grants. Spectrum grant is the process by which a CBSD asks the SAS for authorization to transmit in a particular frequency range at a particular transmission power and follows the Grant state machine shown

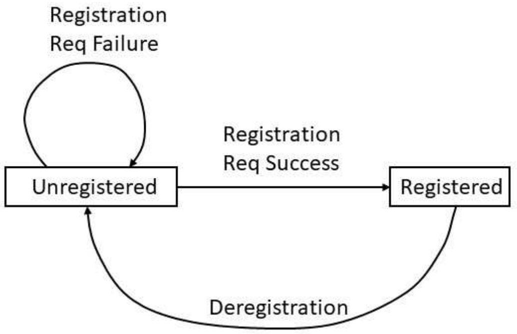


Fig. 1. Registration State Machine of CBRS (adapted from [2])

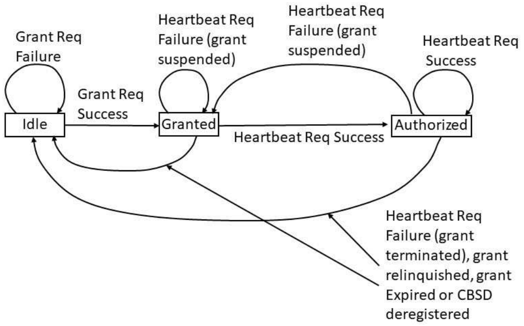


Fig. 2. Grant State Machine of CBRS (adapted from [2])

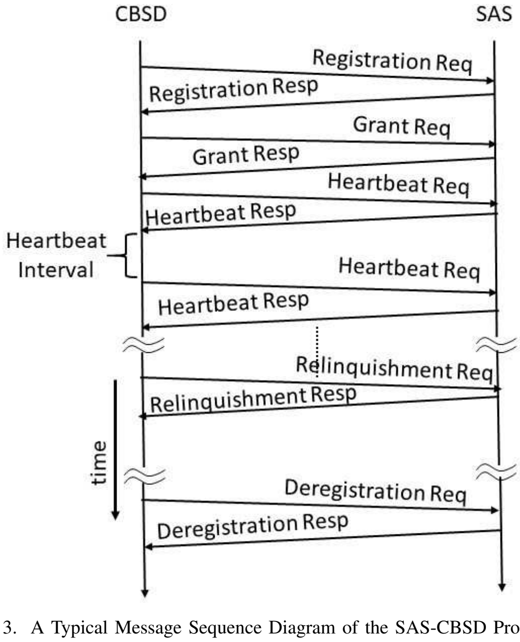


Fig. 3. A Typical Message Sequence Diagram of the SAS-CBSD Protocol

Sahoo, Anirudha; El Ouni, Naceur; Shenoy, Vineet. "A Study of Timing Constraints and SAS Overload in the CBRS Band using SAS-CBSD Protocol." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Wikola, HI, US. December 09, 2019 - December 13, 2019.

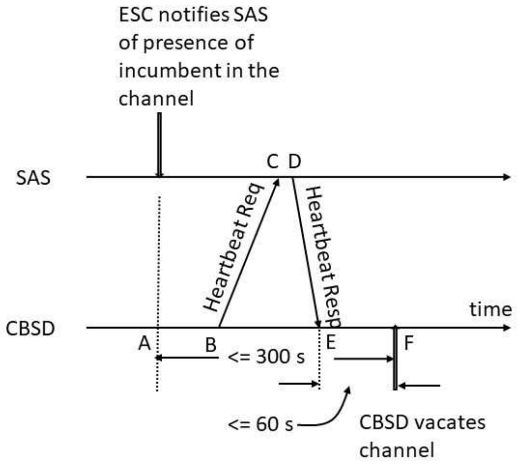


Fig. 4. Timing Constraint of CBRS

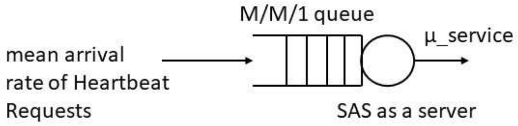


Fig. 5. SAS Message Processing Model

in Figure 2. A CBSD sends a Grant Request to its SAS while in the Idle state. The SAS checks if the grant does not cause harmful interference to other users (as per the Part 96 rules [1]), and if not, the CBSD is granted persmission and it goes into Granted state. If the Grant Request fails, the CBSD stays in the Idle state. A CBSD is not allowed to transmit while in the Granted state. It starts a heartbeat process while in the Granted state by sending a Heartbeat Request message. If the Heartbeat Request is successful, then the CBSD receives a successful Heartbeat Response from the SAS and moves to Authorized state, at which time it is allowed to transmit. A CBSD has to continue sending Heartbeat Requests periodically while in the Authorized state and can continue to transmit as long as it receives successful Heartbeat Responses. The SAS can ask the CBSD to stop transmission by indicating failure in the Heartbeat Response, in which case the CBSD moves to the Granted state from the Authorized state. A SAS may want to do this, for example, if an incumbent appears on the channel used by the CBSD and may experience harmful interference. A CBSD keeps heartbeating in the Granted state to wait for a successful Heartbeat Response from the SAS to move to the Authorized state and resume its transmission. When the CBSD does not want to transmit on that channel any more, it sends a Relinquishment Request to the SAS. The SAS deallocates the resources from the CBSD and sends a Relinquishment Response message to the CBSD. When a CBSD does not wish to participate in the CBRS band any more, it deregisters from its SAS and goes into the Unregistered state as shown in Figure 1. Figure 3 shows a typical successful message sequence diagram of the SAS-CBSD protocol.

Since spectrum is shared in the CBRS band on a priority basis, a SAS may ask a CBSD to vacate its occupied channel to prevent harmful interference to a higher priority incumbent through the Heartbeat Response message. Hence, periodic heartbeat is essential to protect higher priority incumbents. How frequently heartbeats should be sent is decided by the heartbeatInterval parameter which is set in the Heartbeat Response message by the SAS. Since it is possible that Heartbeat Request or Heartbeat Response messages may be lost, there is a transmitExpireTime timer that runs at the CBSD. The value of this timer is decided by the SAS and is carried in a Heartbeat Response message. If Heartbeat Request or Response is lost, then the transmitExpireTime timer will eventually expire at which point the CBSD has to turn off its transmission. There are stringent timing requirements set forth by the FCC for the lower priority user to vacate the channel when a higher priority incumbent radar appears in the same channel. As per the Part 96 Rules [1], a SAS should make sure that when an incumbent appears on a channel, the CBSDs that may cause harmful interference to the incumbent should stop transmitting within $300\,\mathrm{s}$ from the time the SAS is notified about the presence of the incumbent. The FCC rules also specifies that a CBSD has up to $60\,\mathrm{s}$ to turn off its transmission from the time its managing SAS directs it to do so. Thus, the maximum value of transmitExpireTime timer when a CBSD is in the Authorized state is $240\,\mathrm{s}$. Hence, the heartbeatInterval should be less than $240\,\mathrm{s}$ when a CBSD is transmitting in the Authorized state to prevent the CBSD from unnecessarily shutting down its transmitter due to expiry of transmitExpireTime timer. These timing constraints are depicted in Figure 4. A SAS gets notification from the ESC that an incumbent has appeared on the channel at time A. The SAS has to wait for the next Heartbeat Request message from the CBSD to inform it to stop its transmission. The CBSD sends a Heartbeat Request at time B. The SAS sends a Heartbeat Response which carries a command from the SAS to the CBSD to stop transmission. Once the CBSD receives this command at time E it should stop transmission within $60\,\mathrm{s}$. As mentioned earlier, the time the SAS is notified of the presence of an incumbent to the time the CBSD vacates the channel, should not be more than $300\,\mathrm{s}$, i.e., duration between A and F should be less than or equal to $300\,\mathrm{s}$. From the figure it is clear that the heartbeatInterval parameter plays an important role in deciding how soon or late the CBSD will be able to vacate the channel. Note that the worst case happens when SAS gets ESC notification right after it sent out a Heartbeat Response message. In this case, the SAS would get to notify the CBSD to vacate its channel when it gets the next Heartbeat Request that would arrive after heartbeatInterval. Hence, large heartbeatInterval would lead to a SAS taking longer to ensure that a CBSD vacate the channel. In this study, we evaluate performance of CBRS system in terms of time taken to vacate a channel after a SAS is notified of presence of an incumbent radar in the channel.

As mentioned earlier, after receiving the first successful Heartbeat Response, a CBSD goes into the Authorized state and has to continue to heartbeat periodically while it is transmitting in the Authorized state. When there are thousands of authorized grants managed by a SAS, these periodic heartbeats exert significant processing load on the SAS. If processing power of a SAS is not adequately provisioned, then the SAS may not be able to provide Heartbeat Responses in a timely manner. This could lead to transmitExpireTime timer to expire, resulting in unnecessary suspension of CBSDs which do not get Heartbeat Responses in time. Using the SAS-CBSD simulator we evaluate the limits of a SAS in terms of number of CBSDs it can serve successfully.

B. SAS-CBSD Protocol Simulator

1) Message Processing Model: We use the M/M/1 queueing model to represent the message processing service of a SAS provided to various messages sent from its CBSDs (Figure 5). While this model keeps the analysis simple, it gives a fairly good insight into the timing contstraints and SAS overload aspects of the protocol. Among all the messages, the Heartbeat Request message is the most frequent message arriving at the SAS, whereas other messages are relatively infrequent. Hence, the arrival to the M/M/1 queue is approximated by considering only the Heartbeat Request messages. The service rate of the queue is the rate at which the SAS can process a message ($\mu\_service$) and the mean arrival rate (λ_msg_arr) is approximated by the taking the ratio of number of existing grants in the system to the mean heartbeatInterval time. In the simulation, when a message arrives, we check the utilization ($\frac{\lambda\_msg\_arr}{\mu\_service}$) of the queue at that instant, and if it is greater than or equal to 1, then we drop the message. This indicates that the SAS is not provisioned with adequate processing resources and hence the message is dropped.

The above M/M/1 queueing model can be used to calculate the maximum number of CBSDs that a SAS can manage such that there is no unnecessary shut down of CBSDs. Let λ_msg_arr be the mean arrival rate of messages to the SAS, $\mu\_service$ be the mean service rate of the SAS, N_CBSD be the number of CBSDs, G be the maximum number of grants a CBSD is allowed to have and HBI_mean be the mean heartbeatInterval. We consider only Hearbeat Request messages to compute the processing load on a SAS, since other messages arrive relatively infrequently. Hence, the mean message arrival rate at the SAS is given by

$$\lambda\_msg\_arr = \frac{N\_CBSD \times G}{HBI\_mean} \tag{1}$$

To prevent loss of messages at the SAS because the SAS cannot handle the rate at which messages are arriving, the M/M/1 queue should be stable, i.e., utilization of the queue should be less than 1. Hence,

$$\frac{\lambda\_msg\_arr}{\mu\_service} < 1 \tag{2}$$

Using (1) in (2) we get

$$N\_CBSD < \frac{HBI\_mean \cdot \mu\_service}{G} \tag{3}$$

Thus, (3) provides a method to calculate the maximum number of CBSDs a SAS can serve so that it can process messages in a timely manner.

2) Modeling Detection of Incumbent Radar: We have a simple mechanism for modeling detection of incumbent shipborne radar. The incumbent radar appears according to an exponential distribution with mean arrival rate $\lambda\_incumbent$ (see Table I for these parameters and their values) and the channel on which it appears is randomly chosen. When a shipborne radar is detected, we randomly choose $Grants\_affected\_by\_incumbent$ percentage of total grants which are operating in the same channel as the incumbent and put those grants in suspension, i.e., they go into Granted state. We understand that this is far from the real operation of a SAS. In a real operation, a SAS has to compute path loss from each CBSD in the neighborhood of the shipborne radar to the radar receiver and compute the aggregate interference at the radar receiver. If the aggregate interference is more than the specified Interference to Noise (I/N) threshold of -6 dB [4, Requirement R2-IPM-01] then the SAS has to identify which CBSDs should be turned off to bring the aggregate interference down below the threshold. The exact requirement for this operation is specified in [4, Requirement R2-SGN-24]. Even though our shipborne radar protection model does not resemble the real operation, it is adequate to provide essential insight into the time sensitive aspects of vacating a channel by a CBSD.

3) Implementation Details: We have implemented the SAS-CBSD protocol using OMNET++ discrete event simulation software [5]. The core parts of the implementation are the two state machines of the protocol for which we have used the Finite State Machine (FSM) support provided in the Omnet++ simulator [6] and the models presented in the previous sections. The Grant State Machine of a CBSD cannot start unless the Registration State Machine of the CBSD is in the Registered state. Hence the Grant State Machine is implemented as a nested state machine which is supported by Omnet++ [6, Section 4.10.1]. Grant Requests are sent from each CBSD according to a Poisson process, i.e., the interarrival of two consecutive Grant Requests is exponentially distributed. The lifetime of a grant is also exponentially distributed. Each CBSD is allowed to have up to 7 grants. Total number of channels in the system is set to 10. When a CBSD sends a Grant Request, it randomly chooses one of the 10 channels for the grant. The parameters used in our experiments and their values are listed in Table I.

III. Simulations and Results

In this section we describe our experiments and present the results. We primarily ran two types of experiments. The first type is designed to show the effect of overloading the SAS in terms of message handling. In this experiment the focus was only on the message handling of the SAS and hence simulation of presence of incumbent radar was not enabled. The second type of experiment was carried out to study how quickly CBSDs vacate the channel when an incumbent radar appears on the channel to satisfy the timing constraint set forth by the FCC. Table I lists the important parameters and their values used in our experiments.

A. Unnecessary Timeout

In this set of experiments, the mean message service rate of the SAS is fixed (as per M/M/1 queue service rate) while the

Sahoo, Anirudha; El Ouni, Naceur; Shenoy, Vineet. "A Study of Timing Constraints and SAS Overload in the CBRS Band using SAS-CBSD Protocol." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Wikola, HI, US. December 09, 2019 - December 13, 2019.

Parameter	Description	Distribution	value
λ_{grant}	mean interarrival rate of Grant Requests	exponential	$300\,\mathrm{s}^{-1}$
$grant_{life}$	mean lifetime of Grant	exponential	$900\,\mathrm{s}$
λ_{incumb}	mean interarrival rate of incumbent	exponential	$(1/180)\,\mathrm{s}^{-1}$
$incumb_{life}$	mean lifetime of presence of incumbent	exponential	$300\,\mathrm{s}$
$heartbeatInterval$	parameters of heartbeat interval	uniform	[70, 110] s [120. 180] s [200, 240] s
$\mu_{service}$	mean SAS service rate	exponential	40, 60 s^{-1}
T_{sim}	simulation time	–	$86\,400\,\mathrm{s}$ (1 day)
$HB\_success\_rate$	percentage of hearbeat success	–	$100\,\%$
$Grant\_success\_rate$	percentage of grant success	–	$95\,\%$
$Grant\_suspend\_rate$	percentage of grant put into suspension when incumbent appears	–	$100\,\%$
$Grants\_affected\_by\_incumbent$	percentage of existing grant affected when incumbent appears	–	$90\,\%$
G	maximum num of grants per CBSD	–	7
$MAX\_CHANNELS$	maximum number of channels	–	10

TABLE I
PARAMETERS USED IN OUR SIMULATION

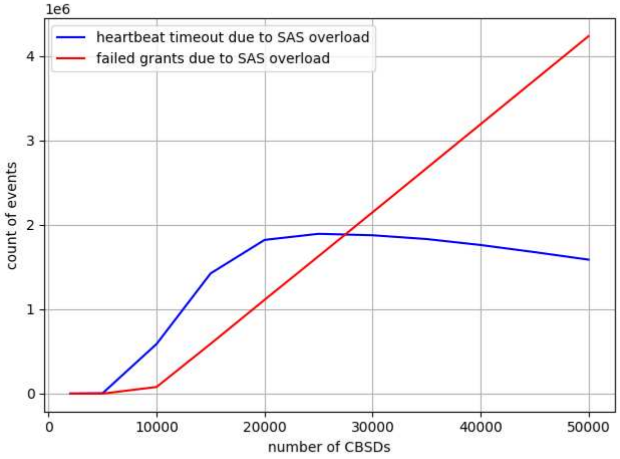


Fig. 6. Unnecessary Timeout vs Num of CBSDs (service rate=40 rps)

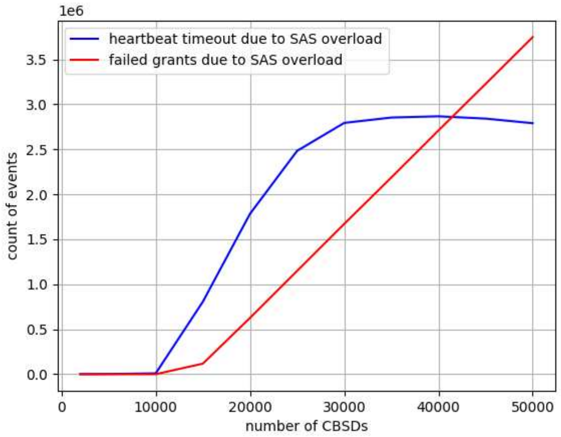


Fig. 7. Unnecessary Timeout vs Num of CBSDs (service rate=60 rps)

number of CBSDs is progressively increased. As the number of CBSDs increases, the number of grants also increases, which in turn increases the number of heartbeat messages to be handled by the SAS. Thus, the message handling load on the SAS increases. At some point the SAS fails to keep up with the rate at which the Heartbeat Request messages arrive. This leads to some Heartbeat Request as well as Grant Request messages being dropped. When a Heartbeat Request message is dropped, the CBSD does not get the corresponding Heartbeat Response message. Hence, the transmitExpireTime timer in the corresponding Grant *unnecessarily* times out. This would force the CBSD to stop its transmission. Clearly this situation arises due to poor provisioning of processing power of the SAS and is not desirable by the commercial operators.

Figure 6 presents the unnecessary heartbeat timeout and failed grants (due to SAS overload) vs number of CBSDs when the mean SAS service rate is 40 requests per second (rps) and the heartbeatIntervals of grants are uniformly distributed between $200\,\mathrm{s}$ and $240\,\mathrm{s}$. For a given number of CBSDs the simulation is run for $86\,400\,\mathrm{s}$ (1 day). The heartbeat timeout remains low until about 5000 CBSDs after which it rises rapidly. But beyond $20\,000$ CBSDs, it tapers off. The failed grant count starts to take off rapidly around $10\,000$ CBSDs and continues to increase. Since grants fail, there are less number of grants in the system. Hence, when the high number of Grant Requests fail, the number of heartbeat failures does not increase and more or less remains flat.

Figure 7 depicts the same performance metrics but with mean SAS service rate of $60\,$rps. The SAS can handle more CBSDs compared to the $40\,$rps case before heartbeats start to time out. For this case the heartbeat fail count takes off at around $10\,000$ CBSDs which is higher than the corresponding point when service rate is 40 rps. However, once the heartbeat fail count rises, it rises more rapidly compared to $40\,$rps case.

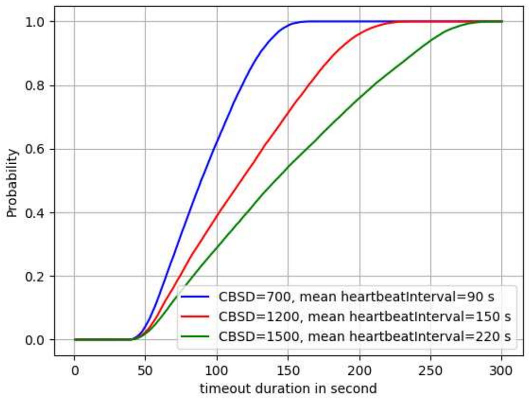


Fig. 8. CDF of Duration of CBSDs Vacating a Channel (service rate=60 rps)

The failed grant count remains zero until about 10 000 CBSDs after which it takes off. The grant fail count happens to be lower than that for the 40 rps case for a given number of CBSDs. Hence, there are more active grants in the system than 40 rps case which implies that there are more heartbeat requests. This leads to a higher heartbeat fail count compared to the 40 rps system.

B. Time to Vacate a Channel

In this experiment, we enabled presence of incumbent radar as per the parameters specified in Table I. When an incumbent radar appears, the SAS randomly chooses 90 % of grants which are in the same channel as the radar and instructs those CBSDs to suspend the grants via Heartbeat Response message. We assume that the SAS sets transmitExpiryTime timer to zero when it commands the CBSD to suspend the grant in the Heartbeat Response message. Hence, the CBSD has up to 60 s to vacate the channel. Figure 8 shows the Cumulative Distribution Function (CDF) of the CBSDs vacating the channel for different mean heartbeatInterval. In this experiment we assume that the SAS is adequately provisioned with processing power such that there is no Heartbeat Request or Grant Request message loss. To achieve this, for a given mean heartbeatInterval, we calculate the number of CBSDs (using (3)) such that the SAS does not drop any message due to overloading of message arrival. For example, when the heartbeatInterval is uniformly distributed between (200, 240) s, the mean hearbeatInterval is 220 s. If a SAS has a service rate of 60 rps, then using (3), the SAS can handle up to 1885 CBSDs. So, we fix the number of CBSDs at 1500. From Figure 8, it can be noticed that when the mean heartbeatInterval increases, more and more grants vacate the channel closer to the 300 s limit. In fact, for a mean heartbeatInterval of 220 s some grants vacate channel very close to 300 s. Since there is randomness involved in the process, it is not advisable to have CBSDs vacating the channel so close to the deadline although it allows more CBSDs to be served for a given SAS service rate. Hence, a heartbeatInterval of 220 s may not be a good choice. On the contrary, if heartbeatInterval is set to a low value of 90 s, then CBSDs can vacate the channel much ahead of the deadline. However, shorter heartbeatInterval exerts heavy load on the SAS which means a smaller number of CBSDs (only 700 CBSDs when mean heartbeatInterval is 90 s) can be served by the SAS for a given service rate. Setting heartbeatInterval to around 150 s seems to be a good choice to achieve a reasonable balance between the number of CBSDs a SAS can manage and the time taken for CBSDs to vacate a channel. Note that the SAS service rate does not have much effect on the CDF of time taken by CBSDs to vacate a channel when SAS is operating close to its full capacity (i.e., close to M/M/1 server utilization of 1). This is true in our study, since the number of CBSDs a SAS manages is set close to the maximum possible value as per (3).

IV. Conclusion and Future Work

We have developed a basic SAS-CBSD protocol simulator to study the impact of heartbeatInterval on the CBRS system in terms of meeting the end to end timing constraint set forth by the FCC rules. We also use the simulator to study how message overload on a SAS leads to unnecessary timeout of transmitExpireTime timer which leads to suspension of CBSD transmission thereby reducing spectrum utilization. Through our experimental results we have shown the tradeoff between time taken to meet the end to end timing constraint and the number CBSDs that can be served by a SAS without causing unnecessary supension of CBSD transmission. With a lower heartbeatInterval the end to end timing constraint of vacating a channel can be met earlier, however, the number of CBSDs a SAS can serve will be lower. Based on our results, setting heartbeatInterval to around 150 s may strike a good balance between the tradeoffs.

Implementing a more practical model to represent the method of vacating a channel when an incumbent shipborne radar appears on the channel can be an useful future work. This would require implementing a propagation model and the so called *move-list* algorithm (which identifies the CBSDs that should vacate the channel when a incumbent shipborne radar appears on the channel) defined in the standards [4].

References

[1] "Citizens broadband radio service," Title 47 Code of Federal Regulations Part 96, 2019.

[2] "Signaling protocols and procedures for citizens broadband radio service (cbrs): Spectrum access system (sas) - citizens broadband radio service device (cbsd) interface technical specification," Wireless Innovation Forum Document WINNF-TS-0016, Version V1.2.4, 2019. [Online]. Available: https://winnf.memberclicks.net/assets/CBRS/WINNF-TS-0016.pdf

[3] M. M. Sohul, M. Yao, T. Yang, and J. H. Reed, "Spectrum access system for the citizen broadband radio service," *IEEE Communications Magazine*, vol. 53, no. 7, pp. 18–25, 2015.

[4] "Requirements for commercial operation in the U.S. 3550–3700 MHz citizens broadband radio service band," Wireless Innovation Forum Document WINNF-TS-0112, Version V1.8.0, 2019. [Online]. Available: https://winnf.memberclicks.net/assets/CBRS/WINNF-TS-0112.pdf

[5] OMNET++ Discrete Event Simulator. [Online]. Available: https://www.omnetpp.org/

[6] OMNET++ Simulation Manual. [Online]. Available: https://www.omnetpp.org/doc/omnetpp/manual/

**Proceedings of the ASME 2019 14th International
Manufacturing Science and Engineering Conference
MSEC2019
June 10-14, 2019, Erie, PA, USA**

MSEC2019-3032

CHALLENGES IN REPRESENTING MANUFACTURING PROCESSES FOR SYSTEMATIC SUSTAINABILITY ASSESSMENTS: WORKSHOP ON JUNE 21, 2018

Arvind Shankar Raman[1], Dustin Harper, and Karl R. Haapala
School of Mechanical, Industrial, and Manufacturing Engineering
Oregon State University
Corvallis, Oregon 97331, USA

Barbara S. Linke
Department of Mechanical and Aerospace Engineering
University of California Davis
Davis, California 95616, USA

William Z. Bernstein and KC Morris
Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland 20899, USA

ABSTRACT

A workshop on Challenges in Representing Manufacturing Processes for Systematic Sustainability Assessments, jointly sponsored by the U.S. National Science Foundation, the U.S. National Institute of Standards and Technology, ASTM International, the American Society of Mechanical Engineers, and the Society of Manufacturing Engineers, was held in College Station, Texas on June 21, 2018. The goals of the workshop were to identify research needs supporting manufacturing process characterization, define limitations in associated education practices, and emphasize on challenges to be pursued by the advanced manufacturing research community. An important aspect surrounded the introduction and development of reusable abstractions of manufacturing processes (RAMP), which are standard representations of unit manufacturing processes to support the development of metrics, methods, and tools for the analysis of manufacturing processes and systems. This paper reports on the workshop activities and findings, which span the improvement of engineering education, the understanding of process physics and the influence of novel materials and manufacturing processes on energy and environmental impacts, and approaches for optimization and decision-making in the design of manufacturing systems. A nominal group technique was used to identify metrics, methods, and tools critical to advanced manufacturing industry as well as highlight the associated research challenges and barriers. Workshop outcomes provide a number of research directions that can be pursued to address the identified challenges and barriers.

Keywords: Unit Process Modeling, Advanced Manufacturing, Nanomanufacturing, Sustainable Manufacturing, Engineering Education

1. INTRODUCTION

The advanced manufacturing research community has been recently exploring wide-ranging issues attendant with additive manufacturing, bio-manufacturing, nanomanufacturing, and smart manufacturing, among other technical domains. However, the integration of findings across these domains, especially in support of sustainable manufacturing analysis, has become increasingly challenging due to the complexity of information and the domain expertise necessary for its interpretation [1]. Further, the reusability of results is inhibited by the lack of standard methods for interpretation, e.g., reproducing results from life cycle assessment (LCA) remains a challenge [2].

The pursuit of sustainable manufacturing requires balancing competing objectives, including cost, time, and environmental and social considerations. The complexity of modeling these objectives increases during the assessment of process-related manufacturing impacts. Assessing system-level sustainability performance is further complicated by the uncertain emergent properties of systems [3], but it is ultimately vital to comparing alternative designs of products and production systems for sustainable manufacturing. Standard approaches for acquiring and exchanging manufacturing process information will lead to more consistent process characterization and may contribute to a consolidated repository of process models for reuse across advanced manufacturing domains [4]. Improved process

[1] Contact author: shankara@oregonstate.edu

modeling (e.g., using data-driven and physics-based approaches), along with the ability to compose a variety of process models, will ensure more effective communication of computational analytics and may facilitate sharing of sustainability performance data [5].

Engaging industrial and academic researchers in presenting and discussing related work is critical for sharing best practices, exposing gaps in technical and educational research and practice, and developing new research agendas. To facilitate this, two workshops have been organized that focused on research to support modeling of manufacturing processes across production scales (e.g., discrete, batch, and continuous) using a variety of manufacturing methods (conventional manufacturing, nanoscale manufacturing, additive manufacturing, and others) applied in various fields (e.g., mechanical, electrical, chemical, nuclear, biochemical, and biological).

This paper presents motivations for the 2018 workshop, based on the prior 2017 workshop (Section 2). It summarizes the workshop activities (Sections 3-6) and presents an outlook for future directions the manufacturing research community might explore (Sections 7). A full workshop report is forthcoming.

2. PRIOR WORKSHOP OUTCOMES

A workshop on Formalizing Manufacturing Processes for Structured Sustainability Assessments, supported by the U.S. National Science Foundation (NSF), was held in conjunction with the 2017 ASME Manufacturing Science and Engineering Conference (MSEC) of the American Society of Mechanical Engineers (ASME) and the 45th North American Manufacturing Research Conference (NAMRC) of the Society of Manufacturing Engineers (SME) on June 7, 2017. The workshop was announced by the National Institute of Standards and Technology (NIST) in partnership with ASTM International, NSF, and ASME. The objectives of the workshop were to:

1. Model different unit manufacturing processes (UMPs);
2. Apply the new ASTM E3012-16 standard for various UMPs; and
3. Provide models suitable for system analysis based on the reusable standard format.

The workshop attracted several dozen participants from industry, academia, and government labs. Results from this workshop [1] highlighted the need for an open repository of process models. The workshop identified many emerging efforts including both standards and research, and outlined a vision for coalescing these efforts towards an open process model repository. In addition, lessons from the 2017 workshop led to proposed revisions of ASTM E3012-16 [6]. Experience through the workshop revealed a need for more rigorous definition of the concepts presented in the standard to support consistent application and implementation. In response to this need, E60.13, the ASTM subcommittee on Sustainable Manufacturing, is revising E3012 with a more robust information model. The forthcoming information model will facilitate more consistent characterizations of physical artifacts in production systems, leading to better reusability of models and reproducibility of environmental analyses.

Based on 2017 workshop results and findings from ongoing research, a 2018 workshop was planned to

1. Provide a venue for participants from industry, government labs, and academia to exhibit manufacturing process developments of their own interest;
2. Identify educational and research challenges and requirements relevant to manufacturing process model development and validation;
3. Expose the research community to developments in the recent standards for modeling manufacturing processes being proposed to the ASTM E60 subcommittee;
4. Identify candidate models to populate an extensible repository of reusable manufacturing process models;
5. Gather inputs on best practices for sharing, reusing, extending, and composing models of conventional and advanced manufacturing processes for characterizing manufacturing systems;
6. Develop a roadmap that defines key research gaps and strategies for addressing system-level modeling; and
7. Enable sharing of model development experiences for evaluating sustainability performance.

3. 2018 WORKSHOP OUTCOMES

The 2018 workshop, titled *Challenges in Representing Manufacturing Processes for Systematic Sustainability Assessments*, supported by NSF's Nanomanufacturing program, was held in conjunction with the 2018 ASME MSEC and the 46th SME NAMRC conferences on June 21, 2018 at Texas A&M University, College Station, TX, and sponsored by ASME, ASTM International, NIST, NSF, and also SME.

The workshop, comprised of two half-day sessions and an evening poster session, engaged the research community in discussions of emerging topics in advanced manufacturing, nanomanufacturing, sustainable manufacturing, and engineering education. The workshop hosted 46 student participants from the NIST RAMP Challenge competition [7], which included six teams of 23 student finalists. Also, there were two dozen participants from industry, academia, and government labs. As part of the workshop, undergraduate and graduate students were able to present their research in manufacturing process development, process modeling, and sustainability performance assessment.

Expected outcomes of the workshop were to identify needs for UMP characterization to support system-level sustainability assessment, to define limitations in associated engineering education and research practices, and to prioritize the challenges to be pursued by the advanced manufacturing research community to best meet industry needs in adopting and applying analytical methods for improving process and system performance. The workshop outcomes summarized in Section 7 were gathered through brainstorming discussions aimed at identifying the barriers and opportunities in several topics of relevance to advanced manufacturing research.

4. 2018 WORKSHOP OVERVIEW

The theme for the 2018 workshop was *Tracking Resources and Flows through the System*. The morning session began with presentations from a NIST-hosted challenge competition on modeling UMPs. After the student presentations, workshop organizers presented use cases of applications exploring unit process modeling and system composition, including those from industry, research, and educational settings, in the form of lightening talks (Section 5). A moderated discussion after the talks fielded audience comments and questions about the cases discussed and points made.

In the afternoon session, facilitated breakout conversations with the attendees were used to gather inputs on methods, issues, and challenges in sharing, reusing, extending, and composing process models for characterizing manufacturing systems, as well as strategies to overcome these challenges, including engineering curriculum development needs. Organizers applied a modified nominal group technique [8], which typically follows this flow: 1) Introduction; 2) Individual idea generation; 3) Idea sharing; 4) Group discussion; and 5) Voting and ranking of ideas. This approach effectively involved all participants, who arrived with varying levels of experience and perspectives on the topics.

Workshop participation was open to MSEC/NAMRC conference attendees with broad interests in teaching engineering students and conducting basic and applied research in manufacturing. Academic researchers with foci in advanced manufacturing, nanomanufacturing, and engineering education were particularly encouraged to attend. NIST RAMP Challenge participants were also encouraged to attend, since they had practical application knowledge based on their work completed for the competition. Section 5 summarizes the workshop lightening talks, which addressed the workshop theme.

5. SYNOPSIS OF THE LIGHTENING TALKS

The lightening talks were presented by workshop organizers and affiliated subject matter experts. Topics focused on various applications of advanced manufacturing technologies in industry and research. The talks centered around the goal of engaging the research community in discussions of emerging topics in advanced manufacturing, nanomanufacturing, sustainable manufacturing, and engineering education.

5.1 Nanomanufacturing

Khershed Cooper, a program director in the NSF Advanced Manufacturing program, defined nanomanufacturing as the fabrication of nano-scale building-blocks (nanomaterials, nanostructures), their assembly into higher-order structures, and the integration of these into larger scale systems with manipulation and control of matter at the nano-scale. Research challenges were noted in processes, metrics, precision, speed of production, unit processes, and integration and packaging. Nanomanufacturing processes should be controllable, reproducible, repeatable, and reliable. Production should be scalable, affordable, safe, have high yields and efficiency. Nano-products should be of high quality, durable, and exhibit desired performance and functionality. With these factors in mind, appropriate metrics to be evaluated can be determined, e.g., precision placement, feature size, and density.

Machine learning has an integral role in nanomanufacturing processes. For example, raw materials serve as an input for an advanced manufacturing system, which outputs meta-materials that exhibit both performance and quality characterized via *in situ* metrology. These characterizations are inputs to a machine learning node. Customer specifications are also an input to the machine learning node. Outputs of the machine learning node are tuned process parameters that feed back into the advanced manufacturing system for process optimization.

Next, Ajay Malshe, a professor of mechanical engineering at the University of Arkansas, outlined three main drivers for standardization in nanomanufacturing: efficiency, yield, and a diverse operating environment. It is important to maintain a business perspective on standardization by keeping factors such as return on investment (ROI) and productive yield in mind. Efficiency is important to business and should be thoroughly considered in developing standard nanomanufacturing methods.

Future nanomanufacturing research efforts are anticipated in three waves: 1) Nanoparticle-based production; 2) Nano-scale template-based production; and 3) True self-assembly for production. Two eminent objectives are 1) Repeatability, reliability, and reproducibility (3Rs), and 2) Product, productivity, and producibility (3Ps). In particular, nano-products should be scalable and minimize waste.

Current nanomanufacturing limitations can be seen through the lens of industry. One limitation is increasing stress levels in the research lab because of a dramatically changing *invention-to-product* life cycle. Further sources of limitations are the complex solutions required. There is a need to account for the frequency of products changing hands. Additionally, there is a missing link between research and industrial application, which could be mitigated by researchers addressing the industry needs. The overall vision for manufacturing science and engineering research is to support the development of the 3Rs and 3Ps for sustainable nano-manufactured products with ROI.

5.2 Systems integration: Additive manufacturing

Kevin Lyons, a senior research engineer in the NIST Systems Integration Division, presented on additive manufacturing and its components: part design using CAD tools, followed by CAD adjustments for additive manufacturing, part build, part post-processing, verification, and validation. Three main focus area in additive manufacturing include industry drivers, research challenges, and scientific and engineering approaches.

Limited connectivity exists between additive manufacturing lifecycle activities and supply chain activities. There is also a disconnect between additive manufacturing software tools, as well as limited process understanding and knowledge of design decision support [9]. The management and representation of additive manufacturing models and knowledge are isolated in industry, and data is generated individually and is costly through additive manufacturing lifecycle activities without coordination. Additionally, because of the heterogeneous nature of these models, it is difficult to combine these models.

The collection and curation of data remains a main research challenge for additive manufacturing. Further, the diversity of the additive manufacturing operating environment gives rise to some important questions, e.g., *How do researchers integrate across the various models while considering the inherent complexities, underlying assumptions, and constraints?* and *How will the models be coordinated?* Design for Additive Manufacturing (DfAM) [10] is an approach to characterize performance and life cycle considerations using design methods or tools for process optimization. Drivers of DfAM include providing manufacturers approaches to capture design rules for different additive manufacturing processes by using formal representations. Additionally, DfAM provides the architecture to derive design rules in a computer-interpretable way, allowing the effective exchange of additive manufacturing information.

5.3 Sustainable engineering education

Fazleena Badurdeen, a professor of mechanical engineering at the University of Kentucky, presented challenges in educating engineers about sustainable manufacturing. There is a need to demonstrate reduced negative environmental impact through sustainable manufacturing, offer improved energy and resource efficiency, provide operational safety, and improve personal health, all at the product, process, and systems levels. Throughout the lifecycle of the product, the 6Rs (reduce, recycle, reuse, recover, redesign, and remanufacture) are implemented at all points, such as redesign during product manufacturing [11].

Realizing sustainable manufacturing innovations requires developing an educated and skilled workforce. This idea falls in line with the United Nation's Sustainable Development Goal 4 (*Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all*) [12]. A lifecycle approach can be applied to recruit, reeducate, and retrain at all levels for building a workforce pipeline [13]. Additionally, a need for a multi-disciplinary approach, which incorporates convergent research and education, to address sustainable manufacturing challenges has to be emphasized. Various programs and funding opportunities can facilitate efforts to bolster sustainable manufacturing education. For example, the NSF Education and Human Resources (EHR) directorate sponsors programs such as the Research on Learning in Formal and Informal Settings. Programs such as these can promote activities that strengthen Science, Technology, Engineering, and Mathematics (STEM) education and prepare future STEM leaders for a rapidly developing work environment.

5.4 Unit Process Life Cycle Inventory (UPLCI)

Barbara Linke, an associate professor of mechanical engineering at the University of California Davis, presented on the Unit Process Lifecycle Inventory (UPLCI) method [14], which is entering its ninth year of development. UPLCI uses industrial information for a single manufacturing process (a machine) to estimate material inputs, energy use, material loss, and dependencies of process parameters with respect to product design. UPLCI is a multi-institutional effort by Katholieke Universiteit Leuven, Northeastern University, Oregon State University, Purdue University, University of California Davis, University of Michigan, University of Virginia, University of Wisconsin, and Wichita State University.

Creating a UPLCI follows a clear, easy-to-follow template, and has breadth to allow different materials and designs. UPLCI require about a month for development, suitable for graduate class project or part of thesis. Moreover, the Springer journal *Production Engineering,* sponsored by the German Academic Society for Production Engineering, now publishes UPLCI studies. The unit process is decomposed into physics-driven equations that describe the energy and material consumption. UPLCIs can be integrated together to evaluate a manufacturing line leading to a completed part or product. The UPLCI approach plans to develop life cycle inventories for around 100 varied manufacturing processes, of which 31 have been completed, with categories including heat treatment, surface finishing, joining, auxiliary, material conversion, and material reducing.

Another unit process inventory approach was created under the Cooperative Effort on Process Emissions in Manufacturing (CO2PE!) framework [15]. Challenges in UPLCI methods include data quality and availability, how to reduce complexity while remaining generic, managing empirical models, materials dependency, energy-dependence on machine set-up, and an unclear vision of whether auxiliary processes are to be included or not. As supporters of UPLCI continue to solve these problems, the demand for inventorying models will continue to increase.

The UPLCI method provides an appropriate means to identify sources of data, collect appropriate models to characterize manufacturing processes, and evaluate performance for general scenarios. However, the storage of UPLCI requires a formal representation to improve robustness. Relating UPLCI studies with ASTM E3012 is a promising approach forward [16].

5.5 Manufacturing process modeling standards

KC Morris, Information Modeling and Testing Group Leader from NIST's Engineering Laboratory and Arvind Shankar Raman, a graduate research assistant in mechanical engineering at Oregon State University, presented a standards-based methodology for extending manufacturing process models for sustainability assessment. They discussed the current lack of assessment tools, which presents analysis challenges. An operational deficiency of analysis applications to support system-, process-, and machine-level manufacturing decisions limits system analysis capabilities. Data collection and reporting has been one of the biggest challenges for manufacturers in pursuing sustainability assessment.

Efforts to characterize manufacturing processes, including UPLCI and CO2PE!, have focused on developing information models that are distinct and specific to processes of concern, sometimes making them limited in their extensibility to related processes. Often these models must be developed from scratch. A question to address becomes, *How do researchers develop methodologies that allow companies to collect, analyze, and disseminate data-driven conclusions about sustainability factors linked to unique manufacturing processes?*

To realize reusability and extensibility, developing models at the appropriate abstraction is critical. Appending an existing model with information about auxiliary systems, such as exhaust gas pressure control systems, monitoring equipment, and electric boosting systems, would constitute a high order variant of the manufacturing process considered [17]. Similarly, if properly developed, models should remain valid after removing information regarding a particular physical setup to be applied to alternative scenarios. For example, in the case of manual drilling, a robust abstraction would facilitate the characterization of multiple instances of drilling regardless of its functional manifestation, e.g., electric hand drilling or using a drill press.

Existing work can be expanded by characterizing data exchange information or linking variables, to facilitate composability. Additionally, an information exchange framework can be created that enables model composability for manufacturing system characterization. Information validation would be critical to its success. The end state of this work would be realizing the framework in commercial software applications.

5.6 Factory Optima: A web-based software

Alex Brodsky, a professor of computer science at George Mason University, presented on a web-based system for composition and analysis of manufacturing service networks based on a reusable model repository [18]. The architectural design allows for rapid software solution development for descriptive, diagnostic, predictive, and prescriptive analytics of dynamic production processes. This architecture emerged in response to the limitations of decision-making tools and models that enable smart manufacturing.

One such limitation is due to the fact that most analysis and optimizations tools are currently developed from scratch, which leads to high cost, long-duration development, and restricted extensibility. Additionally, numerous computational tools are designed to model individual activities, which require the use of specialized, low-level mathematical abstractions. This proposition for an architecture that addresses these limitations is unique in that the middleware layer was based on reusable, modular, and extensible knowledge bases. However, the architecture lacks systematic design of the unit manufacturing process models, which are based on linear functions as opposed to being physics-based, typically involving non-linear functions.

Factory Optima is a high-level system architecture based around a reusable model repository and the unity decision guidance management system (DGMS). The software framework and system enables composition, optimization, and trade-off analysis of manufacturing and contract service networks. This work is unique in its ability to perform tasks on arbitrary service networks without manually crafting optimization models. Industrial case studies are needed to further develop the architecture. In addition, stochastic optimization-based deterministic approximations, and model calibration and training will aid in improving the commercial utility of the tool.

6. BREAKOUT SESSION AND REFLECTION ACTIVITY

Parallel breakout discussions were facilitated by six subject matter experts (the workshop presenters and I.S. Jawahir, a professor of mechanical engineering at University of Kentucky). Discussions were guided by Karl Haapala, an associate professor of manufacturing engineering at Oregon State University, and focused on advanced discrete manufacturing processes, nanomanufacturing at scale, additive manufacturing, process-level sustainability assessment, system-level sustainability assessment, and engineering education in advanced manufacturing. Scribes captured the ideas generated during three timed sessions. These are summarized in Section 6.1.

Each group discussed challenges and opportunities related to metrics and indicators, models and algorithms, and tools and methods. The groups were prompted to progress through the discussion in three four-minute intervals. Participants distributed themselves among the six topic areas and were given 14 minutes per facilitated discussion round to brainstorm ideas related to the topic. The final two minutes of each round were allotted to reviewing the ideas that were shared and collected. The structure of this breakout session allowed for a continuous flow of perspectives and ideas that were guided toward identifying challenges and approaches to overcoming them for each topic.

The final stage of the afternoon workshop session involved an individual reflection activity, which posed two questions: *What do you see as the most pressing need for advanced manufacturing research or advanced manufacturing education?* and *What do you see as the key next step to be taken to address a pressing research or educational challenge in advanced manufacturing?* Participants recorded their individual responses to these questions on notecards, as described in Section 6.2.

6.1 Breakout session results

The breakout session consisted of small-group brainstorming along three subtopic lines: metrics and indicators, models and algorithms, and tools and methods. The results are reported along the same lines. Discussions revolved around challenges, barriers and solutions to overcome them.

Topic 1: Advancing discrete manufacturing processes

Metrics and Indicators: Challenges include product customization, standardization, and bolstering the flexibility of processes. One key barrier is to connect process level controls and system level metrics. Modeling interdisciplinary/dynamic processes can be extremely difficult.

Models and Algorithms: The complexities in model composition and optimization pose barriers to developing flexible models and algorithms. Participants identified a need to support related product categories with similar models across multiple enterprises. Additionally, transient analysis is required for developing robust models of complex systems, especially non-steady state manufacturing elements. Scheduling intricacies pose a challenge for modeling flexible discrete systems.

Tools and Methods: Participants noted that robots, which are widely used in discrete product manufacturing, can be extensively integrated to achieve process improvements. It was also established that machine learning classifications of problems is increasingly an important in advancing the understanding and optimizing the performance of discrete manufacturing processes.

Topic 2: Nanomanufacturing at scale

Metrics and Indicators: Participants identified some of the key metrics and indicators that need to be considered for nanomanufacturing as follows: fluid type, electron beam power, scan rate, beam diameter, material removal rate, structural resolution, feature size, tolerances, nanoparticles (e.g., silver), medium, roll-to-roll, roll speed, printing speed, ink spread, sintering conductivity, circuit device design, and reactor design.

Models and Algorithms: To model the metrics and indicators identified in the above, participants noted existing models and algorithms. Some of the current modeling categories include fluidic modeling, roll-to-roll modeling, circuit modeling, molecular dynamics, and density functional theory (DFT). Participants indicated that currently models or algorithms for other metrics and indicators of interest do not exists.

Tools and Methods: Participants indicated that some of the common tools for modeling and analyzing nanomanufacturing include MATLAB, scanning electron microscopes (SEM), transmission electron microscopes (TEM), computational fluid dynamics (CFD), finite element method (FEM), finite volume method (FVM). The UMP Builder based on ASTM E3012 [19] was also noted as an enabler of analysis. Key advancements in tools have been achieved using machine learning (for prediction), image processing, and fuzzy logic, with advancements in computing technology and an increase in usage of artificial intelligence techniques.

Topic 3: Additive manufacturing at scale

Metrics and Indicators: The participants identified some of the basic metrics for additive manufacturing: temperature, layer thickness, material uniformity, material density, extrusion rates, feed rates, internal geometries, product dimension constraints, melt pool geometries, and build time. More quality-oriented indicators identified were surface profile, accuracy, surface finish, repeatability, preventative maintenance, a need for post-processing operations, and control of multi-axis equipment.

Models and Algorithms: Some of the challenges identified were limitations to support structure optimization, design features, and fidelity of current models. A need exists for topology optimization and an expression of key performance indicators (KPIs) as a function of control parameters. Participants posited cloud-based process design is needed, by combining parameterized product design with process design.

Tools and Methods: The participants desired tools and methods which are able to provide information on selection of process, build orientation, material. Also the tools should be able to support metrology, in-process monitoring tools, quality, verification, validations, sustainable decision support tools, cross-validation tools, selection of models, cost models, and product design optimization methods.

Topic 4: Process level sustainability assessment

Metrics and Indicators: The participants indicated that metrics and indicators for sustainability at the process level include cost, environmental impact, energy, resources, waste minimization, safety, public policy, personal health, productivity and quality, all essentially addressing the three pillars of sustainability. At the process level, these metrics can be difficult to identify and quantify. Safety and public policy, for example, consider societal impacts, legislative and administrative issues, and ethics, which are difficult indicators to effectively assess.

Models and Algorithms: One of the key challenges identified by the participants is the limited models or algorithms that facilitate assessment of process-level sustainability metrics. Physics- and empirical-based methods were discussed, as well as predictive and optimization methods. In addition, participants identified process planning, sensors, and data-driven models as means to assess and improve process-level sustainability.

Tools and Methods: One important topic that emerged as a necessary element of effective sustainability assessment was education. A strong need for bolstering education was identified to address the growing demands and urgency of awareness and accurate sustainability assessment at the process level. Beyond education, the group identified training, skills, societal influence and behaviors as key tools and methods of importance regarding sustainability at the process level.

Topic 5: System level sustainability assessment

Metrics and Indicators: At the system level, lead time and resource availability appear to be metrics and indicators of worthy consideration, as well as material stability, resource availability, and reliability. Additionally, it is important to consider the interaction of multiple manufacturing processes involved at a system level, as one process usually is fed into the next, and the connection between those processes needs to be seamless to ensure more accurate assessment.

Models and Algorithms: For the systems level, it is important that models for risk assessment and for evaluating system dynamics are developed. Models describing manufacturing processes were found to have an important role in system-level sustainability assessment. Also, game theory can be applied iteratively to identify critical issues. Discussions also raised the point that network models should be developed, in addition to unit manufacturing process models.

Tools and Methods: Current challenges for modeling sustainability at the system level include how to collect and sort data. Methods for defining interactions of processes within the system would be helpful. Obtaining a system-level view is essential for the task of sustainability assessment. Participants identified the potential for application of machine learning in predictive modeling of systems level sustainability. Discussions also raised the idea of diagnostic problem identification through degradation classification.

Topic 6: Manufacturing engineering education

Metrics and Indicators: An indicator for education in advanced manufacturing is an identifiable increase in confidence

Shankar Raman, Arvind; Harper, Dustin; Haapala, Karl; Linke, Barbara; Bernstein, William Z.; Morris, KC. "Challenges In Representing Manufacturing Processes for Systematic Sustainability Assessments - Workshop On June 21, 2018." Paper presented at ASME 2019 International Manufacturing Science and Engineering Conference (MSEC2019), Erie, PA, US. June 10, 2019 - June 14, 2019.

in manufacturing classes. Participants also suggested that introducing students to advanced manufacturing at a young age (such as through the use of cartoons) would help increase their interests. A current indicator of weak advanced manufacturing education is the lack of sustainability studies in undergraduate studies. Overall, metrics for engineering education in advanced manufacturing are hard to define.

Models and Algorithms: Some of the models and algorithms associated with engineering education in advanced manufacturing include the applicability of sustainability in real life, easy to apply solutions and methods, and circular design. Additional models taught are design for x (DFx), end of life (EOL), and design for manufacturing (DFM) models (e.g., cost, feasibility, and material use). A robust advanced manufacturing curriculum should include systems engineering models.

Tools and Methods: The tools and methods for bolstering engineering education for advanced manufacturing largely include learning in groups and sharing knowledge. This includes overall manufacturing techniques that can be taught using in-house demonstrations. Basic technical skills to be taught include physics-based classes, which participants suggested being taught in conjunction with case studies and interactive in nature (i.e., labs associated with the material). To provoke students' thinking about sustainability earlier, the group recommended tracking sustainability in real life, and relating sustainability impacts to cost in industry. Hands-on exposure to learning the impacts of manufacturing and relating it to sustainability can be achieved using field trips to manufacturing facilities, for example.

6.2 Results from individual reflection

The answers to the first question (*What do you see as the most pressing need for advanced manufacturing research or advanced manufacturing education?*) were quite diverse, but can be grouped largely into the following categories:

1. Link between research and industry (24%)
2. Development of process models (20%)
3. Improvements in manufacturing education (20%)
4. Advancements in technology and methods of scalability (16%)
5. Encouragement of an interest in manufacturing (12%)
6. Validation of models (8%)

Based on these results, nearly a quarter of the participants thought a stronger link between research and industry was the most pressing need. This indicates a lack of research applications in industry, or at least perception of a lack thereof. The second category, development of process models, scored high as well, likely in response to the workshop discussions tailored toward addressing a need for more models to fill current characterization gaps. Somewhat surprisingly, however, validation of said models did not score as high, even though it was consistently presented as one of the more pressing needs throughout the workshop. This may be a result of an overlap between categories, as some responses qualified for a position in the "link between industry and research" category, but may have also referred to validation.

For the second question (*What do you see as the key next step to be taken to address a pressing research or educational challenge in advanced manufacturing?*), the same six categories were applicable to the responses, in a slightly different order:

1. Improvements in manufacturing education (39%)
2. Link between research and industry (17%)
3. Development of process models (13%)
4. Encouragement of an interest in manufacturing (13%)
5. Advancements in technology and methods of scalability (9%)
6. Validation of models (9%).

As demonstrated here, when posed with questions about the future of manufacturing, a large fraction of the participants regarded education as pivotal to its progression. This and the "encouraging an interest" category are closely related; together they assume a majority of responses to the second question. Based on these results, there is consensus that strengthening the advanced manufacturing community in both numbers and ability is crucial to addressing all the research and industry needs posed during the workshop.

7. FUTURE OUTLOOK

The outcomes of the workshop are expected to benefit research programs, for example, by advancing basic and applied research in topic areas such as sustainability of nanomanufacturing processes and nano-products, digitization of continuous and batch processes, development of physics-based models of manufacturing processes, and efficient process and system models for cloud (cyber) manufacturing. Based on the foregoing, the following research directions emerged:

a) Machine learning methods can support fundamental understanding of a variety of discrete manufacturing processes, e.g., nanomanufacturing, and system-level sustainable manufacturing analysis and optimization.

b) Bridging the gap between process-level controls and system-level metrics can enable deeper insight for discrete and bulk product manufacturing. A mapping of product categories that have similar models and can be used across multiple enterprises is also needed.

c) Transient analysis of complex manufacturing systems can lead to robust manufacturing process models.

d) Metrics and indicators for nanomanufacturing are plentiful and span process parameters, material properties, and part characteristics. They should be unified/harmonized to enable technology comparisons.

e) Scalability in nanomanufacturing needs to lead to reduced defects and defectives, improved metrology, and measurement of moving parts and assemblies.

f) Scalability of additive manufacturing requires material, geometry, and support structure optimization methods.

g) Additive manufacturing key performance indicators must be connected as a function of process controls.

h) In additive manufacturing, integration of *in situ* and out-of-process metrology, sustainability decision tools, model selection tools, cost models, and product design optimization tools, are all areas of research need.

i) Societal influences of sustainable manufacturing, e.g., stakeholder behavior, must be better understood.

j) Engineering education approaches are needed to address the growing urgency for accurate sustainability assessment at the process and system levels.
k) Systemic sustainable manufacturing requires insight from risk assessment and system dynamics methods.
l) Robust methods to characterize interactions of processes, activities, and decisions across a system are needed to advance systemic sustainable manufacturing.
m) Diagnostic problem identification can be aided through degradation classification of physical assets.
n) Developing and sharing metrics for improving the effectiveness of learning in advanced manufacturing should be a focus of engineering education research.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1831141. The authors wish to thank the NSF for their generous support of this research. The authors also extend a warm thank you to the invited speakers, competitors, and supporters of the 2018 workshop.

DISCLAIMER

No endorsement of any commercial product by NIST is intended. Commercial materials are identified in this report to facilitate better understanding. Such identification does not imply endorsement by NIST nor does it imply the materials identified are necessarily the best available for the purpose.

REFERENCES

[1] Garetti, M., and Taisch, M., 2012, “Sustainable Manufacturing: Trends and Research Challenges,” Prod. Plan. Control, 23(2–3), pp. 83–104.

[2] Kuczenski, B., Marvuglia, A., Astudillo, M. F., Ingwersen, W. W., Satterfield, M. B., Evers, D. P., Koffler, C., Navarrete, T., Amor, B., and Laurin, L., 2018, “LCA Capability Roadmap—Product System Model Description and Revision,” Int. J. Life Cycle Assess., 23(8), pp. 1685–1692.

[3] Suh, S., and Qin, Y., 2017, “Pre-Calculated LCIs with Uncertainties Revisited,” Int. J. Life Cycle Assess., 22(5), pp. 827–831.

[4] Bernstein, W. Z., Mani, M., Lyons, K. W., Morris, K. C., and Johansson, B., 2016, “An Open Web-Based Repository for Capturing Manufacturing Process Information,” *Proceedings of the International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, ASME, p. V004T05A028; 8 pages.

[5] Bernstein, W. Z., Bala Subramaniyan, A., Brodsky, A., Garretson, I. C., Haapala, K. R., Libes, D., Morris, K. C., Pan, R., Prabhu, V., Sarkar, A., Shankar Raman, A., and Wu, Z., 2018, “Research Directions for an Open Unit Manufacturing Process Repository: A Collaborative Vision,” Manuf. Lett., 15, pp. 71–75.

[6] ASTM, 2016, “Standard Guide for Characterizing Environmental Aspects of Manufacturing Processes (ASTM E3012).” Conshohocken, Pennsylvania, USA.

[7] Carlisle, M., 2018, “RAMP: Reusable Abstractions of Manufacturing Processes,” NIST [Online]. Available: https://www.nist.gov/news-events/events/2018/01/ramp-reusable-abstractions-manufacturing-processes. [Accessed: 19-Nov-2018].

[8] Stewart, D. W., Shamdasani, P. N., and Rook, D., 2006, *Focus Groups: Theory and Practice*, SAGE Publications, Inc.

[9] Lu, Y., Witherell, P., Lopez, F., and Assouroko, I., 2016, “Digital Solutions for Integrated and Collaborative Additive Manufacturing,” *Volume 1B: 36th Computers and Information in Engineering Conference*, ASME, Charlotte, North Carolina, USA, p. V01BT02A033.

[10] Kim, S., Rosen, D. W., Witherell, P., and Ko, H., 2018, “A Design for Additive Manufacturing Ontology to Support Manufacturability Analysis,” *Volume 2A: 44th Design Automation Conference*, ASME, Quebec City, Quebec, Canada, p. V02AT03A036.

[11] Jawahir, I. S., and Bradley, R., 2016, “Technological Elements of Circular Economy and the Principles of 6R-Based Closed-Loop Material Flow in Sustainable Manufacturing,” Procedia CIRP, 40, pp. 103–108.

[12] “Goal 4: Ensure Inclusive and Equitable Quality Education and Promote Lifelong Learning Opportunities for All — SDG Indicators” [Online]. Available: https://unstats.un.org/sdgs/report/2017/goal-04/. [Accessed: 19-Nov-2018].

[13] Badurdeen, F., and Jawahir, I. S., 2017, “Strategies for Value Creation Through Sustainable Manufacturing,” Procedia Manuf., 8, pp. 20–27.

[14] Overcash, M., and Twomey, J., 2012, “Unit Process Life Cycle Inventory (UPLCI) – A Structured Framework to Complete Product Life Cycle Studies,” *Leveraging Technology for a Sustainable World*, D.A. Dornfeld, and B.S. Linke, eds., Springer Berlin Heidelberg, pp. 1–4.

[15] Kellens, K., Dewulf, W., Overcash, M., Hauschild, M. Z., and Duflou, J. R., 2012, “Methodology for Systematic Analysis and Improvement of Manufacturing Unit Process Life Cycle Inventory (UPLCI) CO2PE! Initiative (Cooperative Effort on Process Emissions in Manufacturing). Part 1: Methodology Description,” Int. J. Life Cycle Assess., 17(1), pp. 69–78.

[16] Brundage, M. P., Lechevalier, D., and Morris, K., 2018, “Towards Standards-Based Generation of Reusable Life Cycle Inventory Data Models for Manufacturing Processes,” J. Manuf. Sci. Eng.

[17] Shankar Raman, A., Haapala, K. R., and Morris, K. C., 2018, “Towards a Standards-Based Methodology for Extending Manufacturing Process Models for Sustainability Assessment,” *ASME 2018 13th International Manufacturing Science and Engineering Conference*, ASME, College Station, Texas, p. V001T05A024.

[18] Brodsky, A., Krishnamoorthy, M., Nachawati, M. O., Bernstein, W. Z., and Menasce, D. A., 2017, “Manufacturing and Contract Service Networks: Composition, Optimization and Tradeoff Analysis Based on a Reusable Repository of Performance Models,” *2017 IEEE International Conference on Big Data (Big Data)*, IEEE, Boston, MA, pp. 1716–1725.

[19] Bernstein, W. Z., Lechevalier, D., and Libes, D., 2018, “UMP Builder: Capturing and Exchanging Manufacturing Models for Sustainability,” *Volume 1: Additive Manufacturing; Bio and Sustainable Manufacturing*, ASME, College Station, Texas, USA, p. V001T05A022.

Proceedings of the
Annual Stability Conference
Structural Stability Research Council
St. Louis, Missouri, April 2-5, 2019

Evaluating Critical Temperatures of Axially Loaded I-shaped Steel Members Using ANSI/AISC-360 Appendix 4 and Finite Element Model

Ana Sauca[1], Chao Zhang[2], Mina Seif[3], Lisa Choe[4]

Abstract

Stability is paramount to the load-carrying capacity of structural steel members subjected to fire. Actual buckling strengths of steel members in fire become lower than that at ambient temperature since modulus of elasticity and yield strength significantly diminish with increasing temperatures. Appendix 4 of the ANSI/AISC-360 specification provides an equation for calculating the flexural buckling stress of columns at temperatures greater than 200 °C. However, finite element analysis showed that columns can fail by buckling at temperatures below 200 °C if the applied axial load is greater than 80 % of its ambient compressive strength. This paper presents: (i) critical temperatures estimated using both the Appendix 4 equations and finite-element analysis and (ii) parametric study results showing effects of applied load level, member slenderness, and steel grades on the critical buckling temperature. Closed form equations are developed and presented along with limitations that were identified.

1. Introduction

The critical temperature method has been considered a useful tool to evaluate failure temperatures of loaded steel members exposed to a standard fire (e.g., ISO-834 (International Standard ISO 834-1:1999)). The Eurocode 3 (EC3) standard (EN 1993-1-2, 2005) specifies the critical temperature method for steel members supporting given axial or flexural loads. This method assumes that a loaded steel member is heated uniformly along its length and across the cross-section. For steel members without any instability phenomenon (e.g., tensile or flexural members) the critical temperature ($\theta_{a,cr}$) can be determined using Eq. (1) which is a function of the degree of utilization (μ_0) in Eq. (2), where ($E_{fi,d}$) is the design effect of actions for the fire design situation; ($R_{fi,d,0}$) is the corresponding design resistance of the steel member for the fire design situation at time (t) = 0.

[1] Guest Researcher, National Institute of Standards and Technology (NIST), <ana.sauca@nist.gov>
[2] Guest Researcher, National Institute of Standards and Technology (NIST), <chao.zhang@nist.gov>
[3] Research Structural Engineer, National Institute of Standards and Technology (NIST), <mina.seif@nist.gov>
[4] Research Structural Engineer, National Institute of Standards and Technology (NIST), <lisa.choe@nist.gov>

Sauca, Ana; Zhang, Chao; Seif, Mina; Choe, Lisa. "Evaluating Critical Temperatures of Axially Loaded I-shaped Steel Members Using ANSI/AISC-360 Appendix 4 and Finite Element Model." Paper presented at Annual Stability Conference, St. Louis, MO, US. April 02, 2019 - April 05, 2019.

$$\theta_{a,cr} = 39.19ln\left[\frac{1}{0.9674{\mu_0}^{3.833}} - 1\right] + 482 \tag{1}$$

$$\mu_0 = \frac{E_{fi,d}}{R_{fi,d,0}} \tag{2}$$

For steel members prone to instabilities (e.g. columns), the critical temperature is presented in specific tabulated data (Vassart et al., 2014), depending upon the degree of utilization, the steel grade, and the non-dimensional slenderness. The non-dimensional slenderness can be determined using Eq. (3), where (*A*) is the gross cross-sectional area (for class 1, 2 and 3 cross sections) or the effective area of a cross section (for class 4 cross sections), (F_y) is the yield stress and (N_{cr}) is the elastic critical force for the relevant buckling mode based on the gross cross-sectional properties. In Eurocode 3 (EC3), for class 4 cross sections, the resistance and rotation capacity is limited by their local buckling resistance.

$$\bar{\lambda} = \sqrt{\frac{AF_y}{N_{cr}}} \tag{3}$$

Some national annexes of Eurocodes specify values for critical temperatures. For example, the British Standard (BS NA EN1993-1-2, 2005) provides critical temperatures for columns and beams. When evaluating the beams, the values of specified critical temperatures vary with fire protection, the support conditions of floor slabs, and the degree of utilization. When evaluating columns, the values of specified critical temperatures vary with the non-dimensional slenderness and the degree of utilization.

Appendix 4 of the American Institute of Steel Construction (AISC) specification for structural steel buildings, known as ANSI/AISC-360 (AISC, 2017), provides advanced and simple methods of analyses for fire conditions; however, it omits the critical temperature method. Hence, the objective of this study is to evaluate and compare the critical temperature of axially loaded I-shaped steel members using the simple method of analysis specified in ANSI/AISC-360 Appendix 4 and finite-element models. The parameters influencing critical temperatures were evaluated, such as the effects of applied load levels, member slenderness ratios, steel grades, and section compactness. Closed-form equations were developed and presented along with limitations that were identified.

2. Critical temperature of steel members using ANSI/AISC-360 Appendix 4

Buckling strengths of steel members subject to fire conditions decrease as the modulus of elasticity (*E*) and yield strength (F_y) diminish with increasing temperatures. Appendix 4 of ANSI/AISC-360 specifies Eq. (4) for calculating the flexural buckling stress (F_{cr}) at temperatures greater than 200 °C, where ($F_y(T)$) is the yield stress at elevated temperature; ($F_e(T)$) is the critical elastic buckling stress calculated from Eq. (5); (*E(T)*) is the elastic modulus at elevated temperatures; (L_c) is the effective length of member; and (*r*) is radius of gyration which is equal to $(I/A)^{0.5}$, where (*I*) is the area moment of inertia and (*A*) is the cross-sectional area.

$$F_{cr}(T) = \left[0.42^{\sqrt{\frac{F_y(T)}{F_e(T)}}}\right] F_y(T) \tag{4}$$

$$F_e(T) = \frac{\pi^2 E(T)}{\left(\frac{L_c}{r}\right)^2} \tag{5}$$

Chapter E of ANSI/AISC-360 uses Eq. (6) and Eq. (7) to calculate the flexural buckling stress (*F_{cr}*) at ambient conditions depending on the value of (L_c/r). In these equations, the yield stress (F_y), the Young's modulus (E), and the critical elastic stress (F_e) are all ambient temperatures values.

When $\frac{L_c}{r} \leq 4.71\sqrt{\frac{E}{F_y}}$ (or $\frac{F_y}{F_e} \leq 2.25$)

$$F_{cr} = \left(0.658^{\frac{F_y}{F_e}}\right) F_y \tag{6}$$

When $\frac{L_c}{r} > 4.71\sqrt{\frac{E}{F_y}}$ (or $\frac{F_y}{F_e} > 2.25$)

$$F_{cr} = 0.877 F_e \tag{7}$$

For columns with compact sections, the nominal compressive strength for flexural buckling at ambient conditions is computed using the flexural buckling stress at ambient conditions (F_{cr}) and the gross cross-sectional area (A) as presented in Eq. (8). For elevated temperatures, Eq. (4) from Appendix 4 of the ANSI/AISC-360 replaces Eq. (6) and (7) to calculate the nominal compressive strength for flexural buckling.

$$P_n = F_{cr} A \tag{8}$$

For columns with slender sections, the nominal compressive strengths at ambient conditions is computed using Eq. (9), where (A_e) is the effective areas of the cross section.

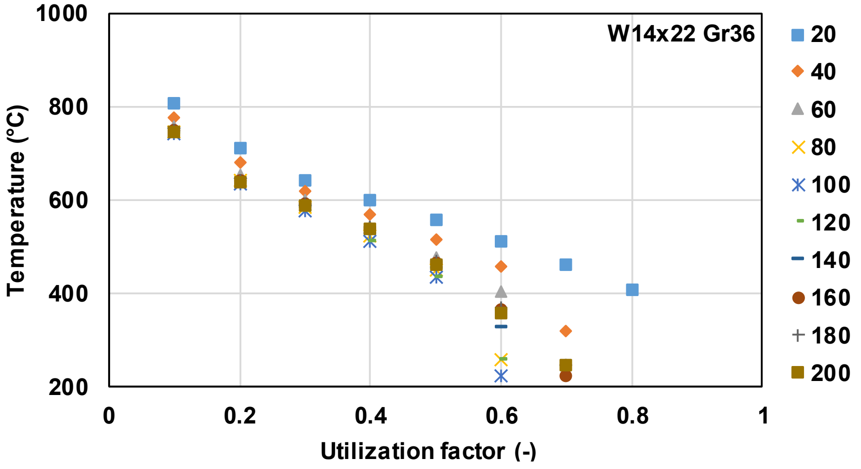
$$P_n = F_{cr} A_e \tag{9}$$

In this paper, the critical temperature can be calculated using the code equations presented above for the range of parameters presented in Table 1. The code-specified minimum yield stress (F_y) and modulus of elasticity of steel (E) are considered in this study. The uncertainty associated with the actual mechanical properties is not considered in this study. The utilization factor is defined as the ratio of $F_{cr}(T)$ in Eq. (4) to F_{cr} in Eq. (6) and Eq. (7) regardless of the section width-to-thickness ratios. The columns are considered simply supported.

Table 1. Considered data to evaluate the critical temperature

Section	Steel	Slenderness $\frac{L_c}{r}$	Utilization factor $\frac{F_{cr}(T)}{F_{cr}}$
W 14x22	Grade 36 (F_y=250 MPa)	20 – 200	0.1 – 1
	Grade 50 (F_y=345 MPa)	20 – 200	0.1 – 1
W 14x90	Grade 36 (F_y=250 MPa)	20 - 200	0.1 - 1
	Grade 50 (F_y=345 MPa)	20 - 200	0.1 - 1

Figure 1 presents the calculated critical temperatures of steel columns using ANSI/AISC-360 Appendix 4 as a function of the utilization ratio for various slenderness values. For columns with utilization factors smaller than 0.5, the effect of slenderness on the critical temperature is deemed small. At the same load level (i.e., utilization factor), the standard deviation of critical temperatures for all slenderness levels was less than 5 % of the averaged value. For utilization factors greater than 0.6, the critical temperature varies significantly with the column slenderness levels. Furthermore, for columns with slenderness values greater than 60 (except slenderness values of 160 and 200) and utilization factors greater than 0.6, the critical temperatures fall below 200 °C, which violates the temperature limit specified for use of Eq. (4) and Eq. (5). Hence, there is a need to limit the utilization factor and slenderness to compute critical temperatures using the AISC equations. The steel grade also has a minor impact on the critical temperatures for most of slenderness levels. Some variations in critical temperatures are observed for shorter columns (with slenderness values less than 40) at utilization factors higher than 0.6.



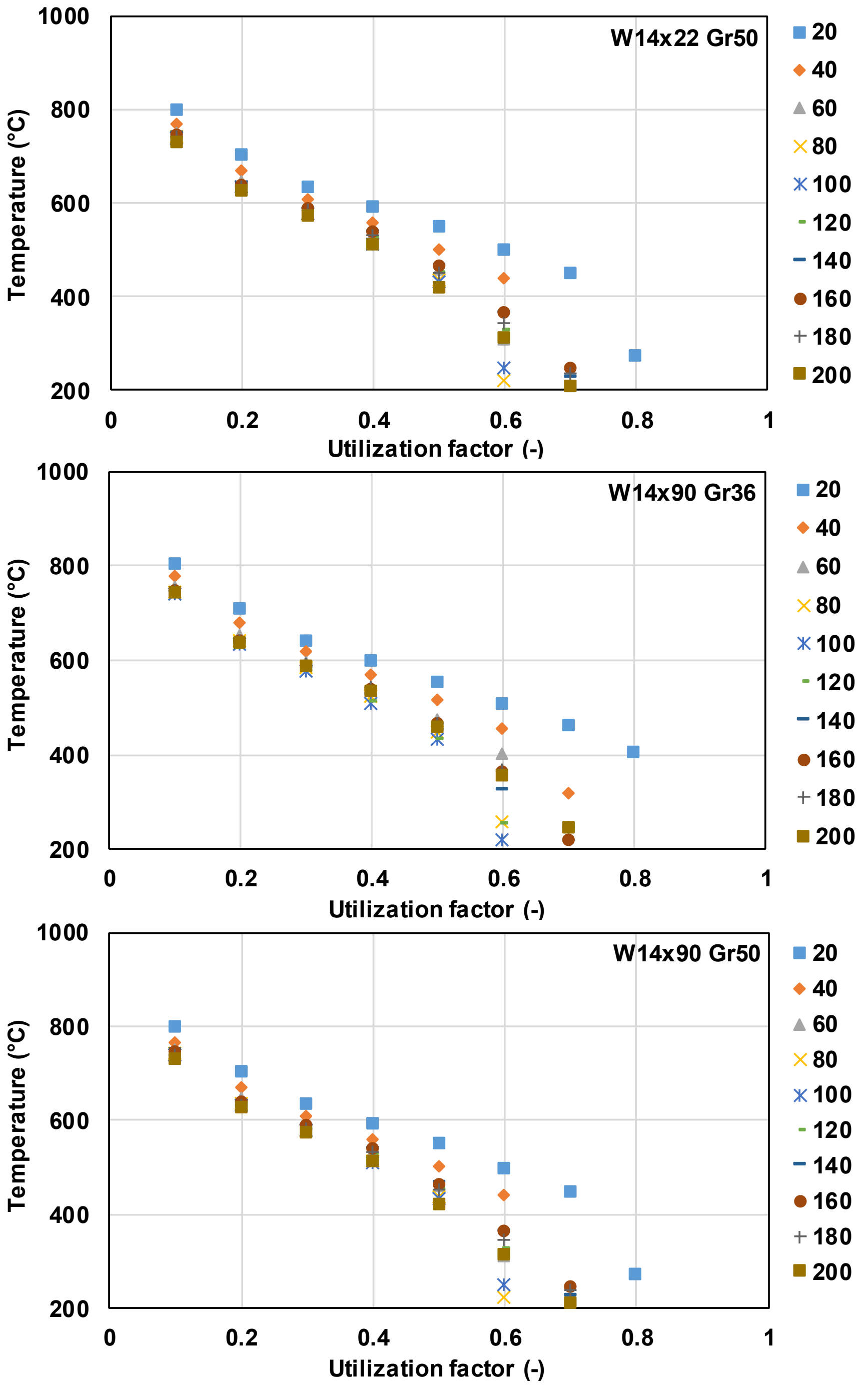


Figure 1. Critical temperature values computed using ANSI/AISC-360 Appendix 4 versus the utilization factor for various slenderness values

Sauca, Ana; Zhang, Chao; Seif, Mina; Choe, Lisa. "Evaluating Critical Temperatures of Axially Loaded I-shaped Steel Members Using ANSI/AISC-360 Appendix 4 and Finite Element Model." Paper presented at Annual Stability Conference, St. Louis, MO, US. April 02, 2019 - April 05, 2019.

3. Critical temperature of steel members using finite element analysis

The cases presented in Table 1 were further evaluated using the finite element analysis software package ANSYS (ANSYS, 2012). The steel column models were meshed with shell element SHELL181, which is a 4-node element with six degrees of freedom at each node. SHELL181 was used because it is suitable for linear, large rotation, and/or large strain nonlinear applications (Index of Software Ansys, 2018), and the change in shell thickness can be accounted for in the nonlinear analysis. The stub and slender columns were meshed using 50 elements for the length, 8 elements for the flange and 8 elements for the web based on the mesh density study presented in Table 2. Linear kinematic constraints were applied to the flanges and web at the column ends to enforce "rigid" planar behavior. At the column ends, the supports were simply supported. Axial force was applied on the center node of the end sections. Global and local initial geometric imperfections were implemented. The initial displacement at midspan was taken as the 1/1000 of the column length to simulate global geometrical imperfections. Local geometrical imperfections were implemented by scaling the deformation of the first (lowest) eigenmode from buckling analysis. The scaled value was the larger of a web out of flatness of (*d/150*) (Kim and Lee, 2002) or a tilt in the compression flanges of (b_f*/150*) (Zhang et al., 2015), where (*d*) and (b_f) are the height and width of the cross section, respectively. It is noted that for short columns, the local buckling modes are expected to dominate the stability behavior. No residual stresses were applied in the numerical models since the effect is regarded to be rather limited in structural fire analysis (Vila Real et al., 2007). A uniform temperature distribution was assumed through the column cross section and along its length. The EC3 material model was used in this study.

The finite element model of the columns is shown in the Figure 2. Some results of the uncertainty component associated with the size of mesh elements (Table 2) as well as the time step associated with the convergence criteria employed in numerical analysis (Table 3) are presented. In a mesh density study, When the column length, flange, and web consisted of the number of elements greater than 100, 16, and 16, respectively, the computed results were converged (0% difference). For computational efficiency, this study used 50 elements for the column span, 8 elements for each flange, and 8 elements for web. This yielded about 2 % difference for the slenderness ranging from 20 to 160. However, the difference in the time step did not significantly the convergence of predicted results. As shown in Table 3, both the increment of the analysis time and duration of the analysis time yields less than 1% difference.

Table 2. Mesh density study

	% difference for W14x22 Gr36		
Number of mesh elements (Column Length x Flange x Web)	**Slenderness 20**	**Slenderness 100**	**Slenderness 160**
100 x 16 x 16	-	-	-
50 x 8 x 8 (used in this study)	-0.20	-1.87	-1.57
30 x 4 x 4	-0.79	-3.53	-9.06

Table 3. Time step influence on the results

	% difference for W14x22 Gr36
Time step (Initial - Min - Max)	**Slenderness 100**
10 - 0.001 - 100	-0.42
1 - 0.01 – 100 (used in this study)	-0.42
0.1 - 0.01 - 10	-0.42
0.1 - 0.001 - 1	-

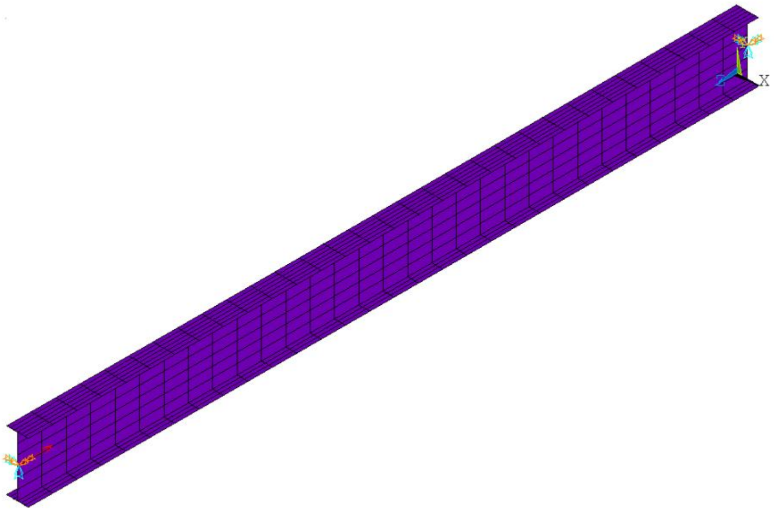
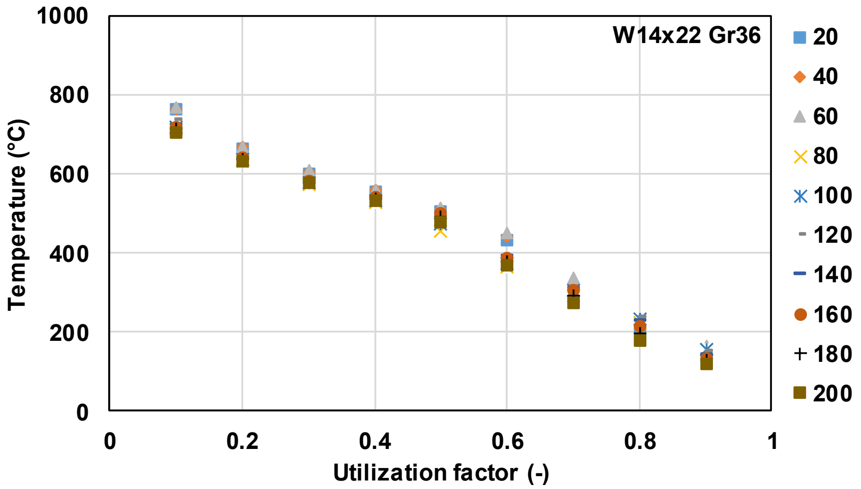
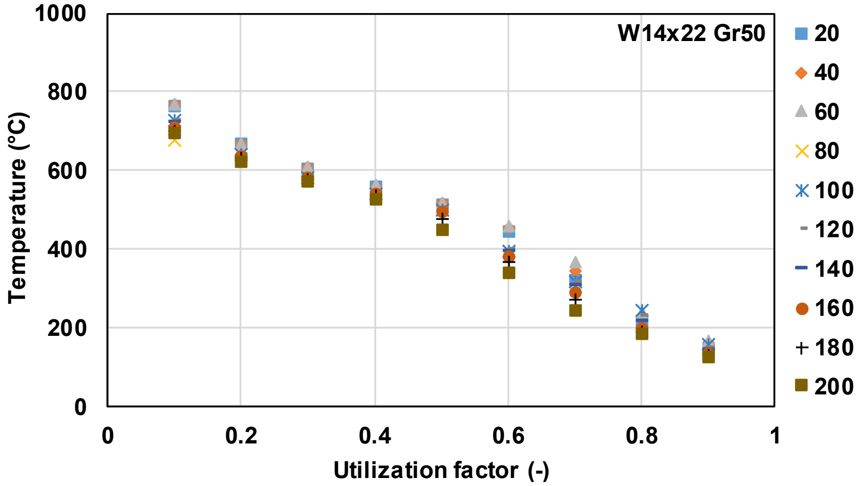
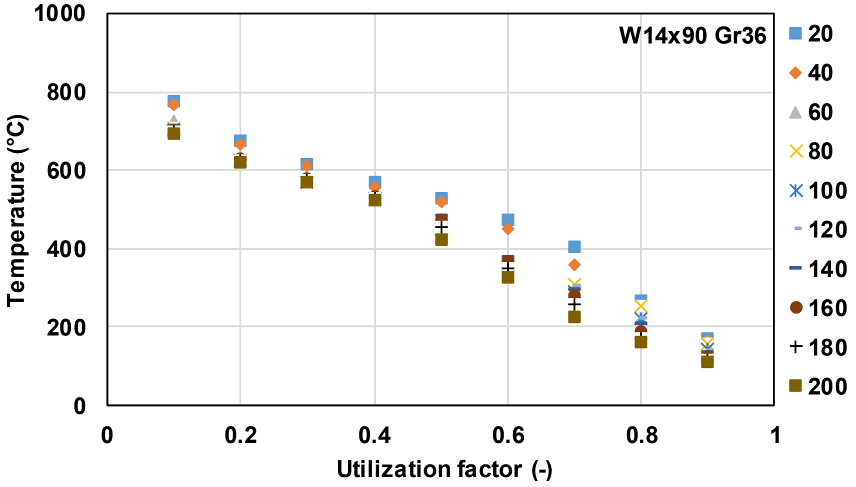


Figure 2. Shell finite element mesh and boundary conditions (ANSYS)

Figure 3 presents the evaluated critical temperatures of steel members using ANSYS versus the utilization ratio for various slenderness values. The critical temperature decreases almost linearly with increasing utilization factors for all slenderness levels. For W14x22 columns, the slenderness has a minor impact on the critical temperature at the same utilization factor. For W14x90 columns, a larger scatter was observed in critical temperatures at utilization factors greater than 0.4.


W14x22 Gr36
Temperature (°C)
0
200
400
600
800
1000
0
0.2
0.4
0.6
0.8
1
Utilization factor (-)
■ 20
◆ 40
▲ 60
× 80
✻ 100
- 120
− 140
● 160
+ 180
■ 200


W14x22 Gr50
Temperature (°C)
0
200
400
600
800
1000
0
0.2
0.4
0.6
0.8
1
Utilization factor (-)
■ 20
◆ 40
▲ 60
× 80
✻ 100
- 120
− 140
● 160
+ 180
■ 200


W14x90 Gr36
Temperature (°C)
0
200
400
600
800
1000
0
0.2
0.4
0.6
0.8
1
Utilization factor (-)
■ 20
◆ 40
▲ 60
× 80
✻ 100
- 120
− 140
● 160
+ 180
■ 200

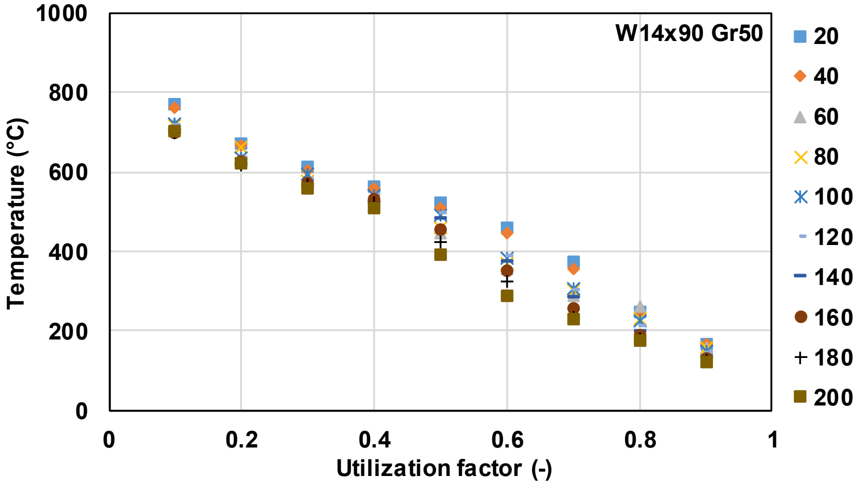
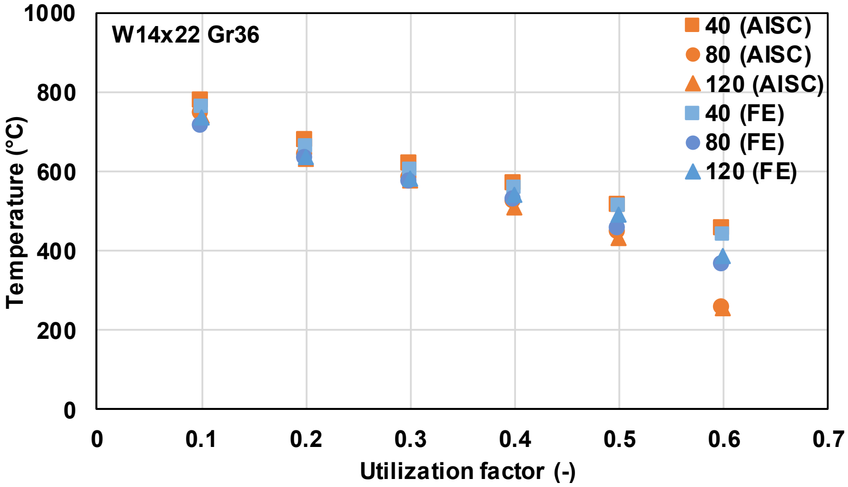


Figure 3. Critical temperature versus the utilization factor for various slenderness values using ANSYS

4. Critical temperatures of steel members using ANSI/AISC-360 Appendix 4 versus finite element analysis

Figure 4 shows a comparison of the computed critical temperatures using ANSI/AISC-360 Appendix 4 versus those computed using ANSYS at practical ranges of utilization factors (0.1 to 0.6) (Moynihan and Allwood, 2014) for slenderness values of 40, 80 and 120. The largest difference between the two methods is as large as 50% for the utilization factor of 0.6. In all the other cases, the temperatures estimated using these two methods are similar (with differences less than 10 %). The standard uncertainties of the finite element model based on the mesh density study and time step influence on the results are around 2.5%.



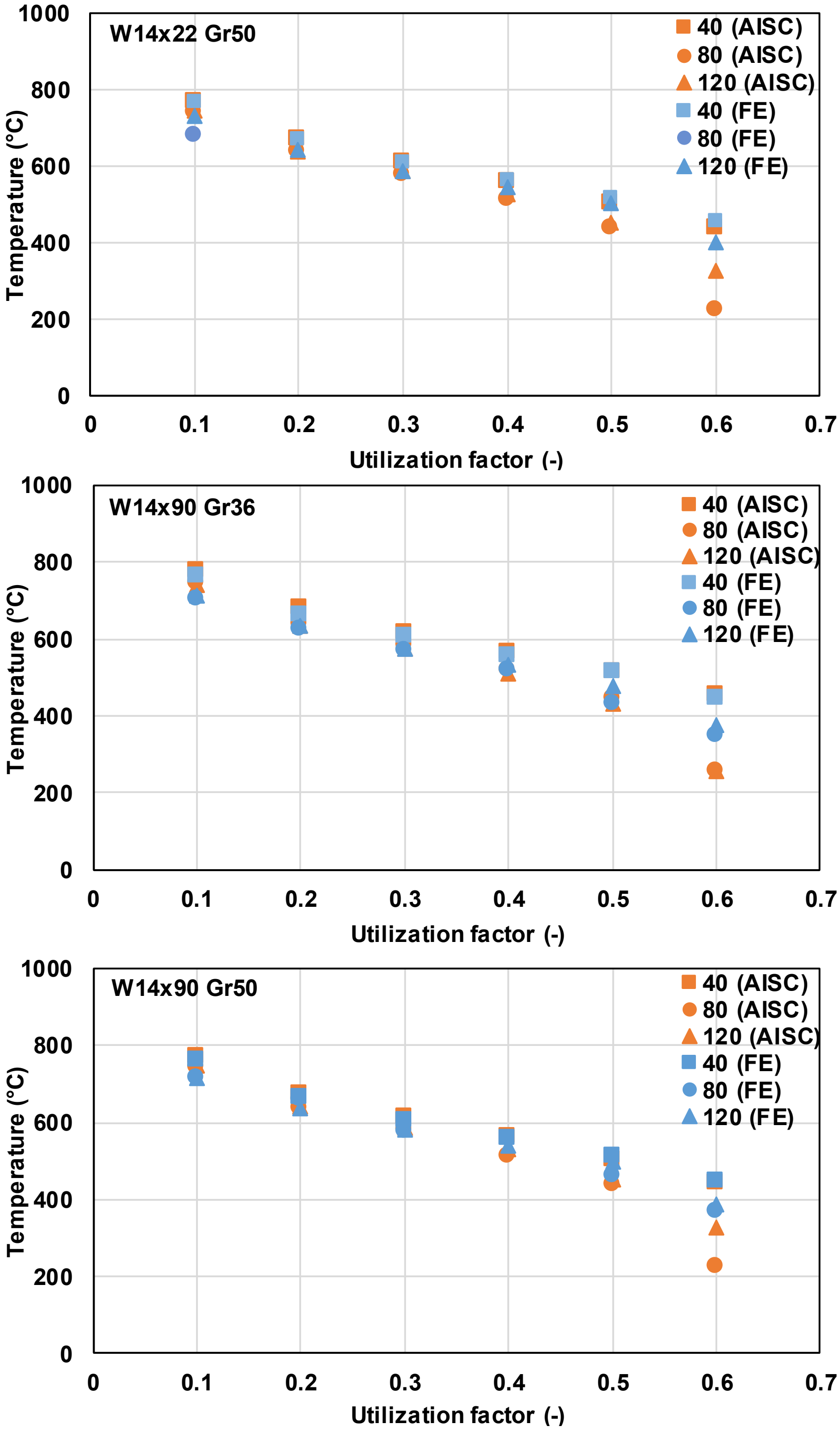


Figure 4: Comparison of critical temperatures versus the utilization factor for slenderness of 40, 80 and 120 based on AISC and ANSYS results

5. Closed formed equation for critical temperature

A closed formed equation for evaluating the critical temperature of steel columns was developed based on the results from the finite element analysis. The considered data were the ones with utilization factors less than 0.6 and temperatures above 400 °C. That is because in practice, (i) the utilization ratio of the columns is between 0.2 and 0.6 (Moynihan, M., C., and Allwood, 2014), and (ii) the degradation of the yield strength of the steel at elevated temperatures starts at 400 °C (Table A-4.2.1 of ANSI/AISC-360).

The critical temperature of steel columns is described in Eq. (10) and is a linear logarithmic function dependent upon the slenderness and the utilization ratio of the columns. The critical temperature equation was determined following the principles presented in the NIST/SEMATECH e-Handbook of Statistical Methods (http://www.itl.nist.gov/div898/handbook/). The standard error, the correlation coefficient (r) and the coefficient of determinations (r^2) are 21.20, 0.97 and 0.95 respectively.

The difference between the critical temperatures estimated using Eq. (10) and the finite element model is below 5% for utilization factors up to 0.4. The maximum difference of 14% occurred at a utilization factor of 0.5 and slenderness of 80.

$$T(°C) = 465.1 + 16.7ln\frac{r}{L_c} + 150.9ln\frac{F_{cr}}{F_{cr}(T)} \tag{10}$$

Figure 5 presents the 3D curve fitting of the data.

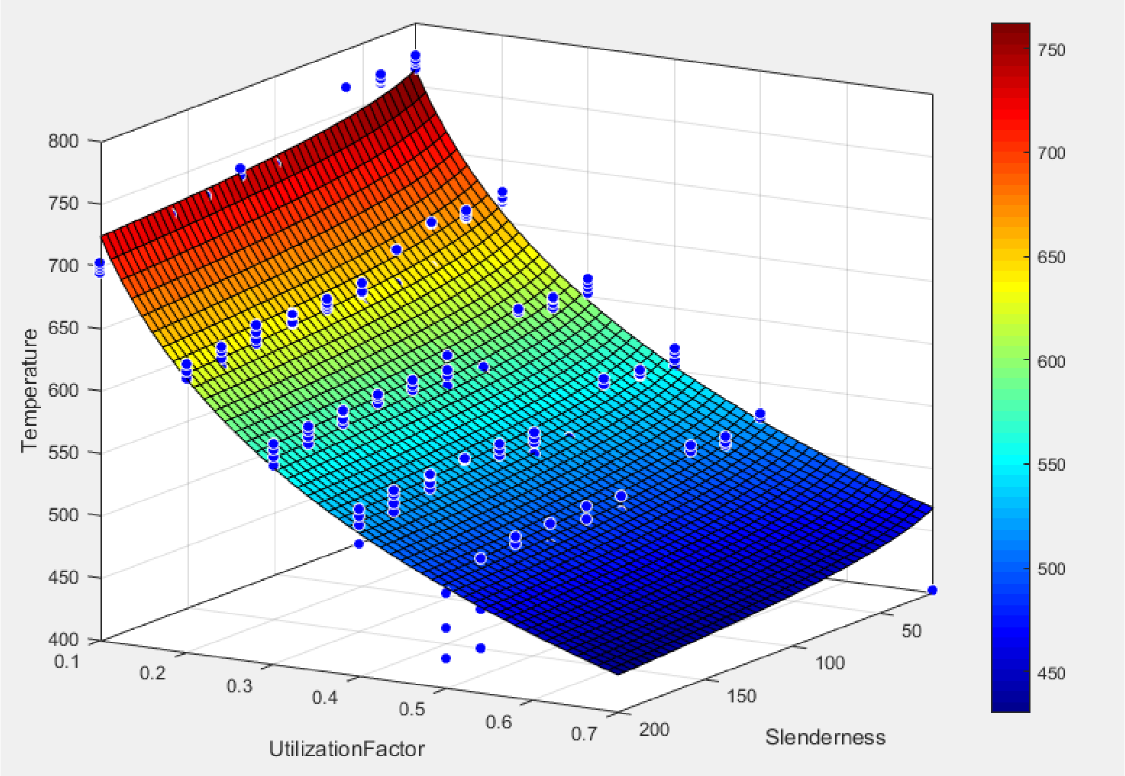


Figure 5: Curve fitting of the data

The proposed equation assumes that all columns cross-sections are uniformly heated, and the boundary conditions are simply supported. The proposed equation was compared with the British National Annex and the results showed differences under 10 %. Ongoing work expands this study, and includes a larger variety of cross-sections, material properties, and boundary conditions.

6. Summary and conclusions

Critical temperatures of steel columns were evaluated using both ANSI/AISC-360 Appendix 4 equations and ANSYS finite element analysis. Both methods yielded similar critical temperatures for utilization ratios less than 0.6. The parametric study shows that for utilization ratios less than 0.6, the impact of member slenderness and steel grades on the critical temperature is deemed small. For cases with a degree of utilization of 0.6 or higher, however, it is recommended to use finite element analysis. A closed form equation presented in this paper can also be used to estimate the critical temperatures of simply supported steel columns as a function of the member slenderness and the utilization ratio (below 0.6). Future work is needed to evaluate a larger range of section types, of material properties and boundary conditions.

Disclaimer

Certain commercial entities, equipment, products, software, or materials are identified in this paper in order to describe a procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, products, software, materials, or equipment are necessarily the best available for the purpose.

References

ANSI/AISC-360 (2017). “Steel Construction Manual”. *American Institute of Steel Construction*

ANSYS user manual, version 14.0 ANSYS Inc. (2012)

ASTM E119-18. “Standard Test Methods for Fire Tests of Building Construction Materials”. *American National Standard*, West Conshohocken, PA

BS NA EN 1993-1-2 (2005): UK National Annex to Eurocode 3. Design of steel structures. General rules. Structural fire design

EN 1993-1-2 (2005): Eurocode 3: Design of steel structures - Part 1-2: General rules - Structural fire design [Authority: The European Union per Regulation 305/2011, Directive 98/34/EC, Directive 2004/18/EC]

Hyams, D.G. (2018). CurveExpert Professional Documentation. Release 2.6.5

Index of Software Ansys. (Last modified May 2018). Retrieved from https://www.sharcnet.ca/Software/Ansys/16.2.3/en-us/help/ans_elem/Hlp_E_SHELL181.html

International Standard ISO 834-1:1999, Fire Resistance Tests—Elements of Building Construction—Part 1: General Requirements.

Kim S., Lee D. (2002). “Second-order distributed plasticity analysis of space steel frames”. *Engineering Structures*, 24 (2002), pp. 735-744

Moynihan, M., C., and Allwood, J., M. (2014). “Utilization of structural steel in buildings”. Proc. R. Soc. A 470: 20140170. http://dx.doi.org/10.1098/rspa.2014.0170

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, January 31, 2019.

Vassart, O., Zhao, B., Cajot, L.G., Robert, F., Meyer, U., Frangi, A. (2014). “Eurocodes: Background & Applications Structural Fire Design”. *JRC Science and Policy Reports*

Vila Real, P.M.M., Lopes da Silva, N.L.S., Franssen, J.M. (2007). “Parametric analysis of the lateral–torsional buckling resistance of steel beams in case of fire”. *Fire Safety Journal*, 42 (2007) [461-24]

Zhang, C., Choe, L., Seif, M., Zhang, Z. (2015). “Behavior of axially loaded steel short columns subjected to a localized fire”. *Journal of Constructional Steel Research*, 2015; 111:103-111.

Dynamic Compressive Tests of Alumina Dumbbells Using a Spherical Joint

Steven Mates[1], Richard Rhorer[2] and George Quinn[1]
[1]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD, USA
[2]Rhorer Precision Engineering, Gaithersburg, MD, USA

ABSTRACT

The dynamic compressive strength of ceramic armor materials is difficult to obtain experimentally due to the sensitivity of fracture strength to even small misalignments and end and surface effects. In this work we introduce a spherical joint into a compression Kolsky bar to investigate whether the joint can alleviate bending stresses due to minor misalignment in compression tests on alumina dumbbell specimens. Tests are conducted both with and without the spherical joint, and high-speed (75 000 frames/s) three-dimensional Digital Image Correlation (3D DIC) is used to measure both the strain field on the dumbbell specimen and motion, if any, in the spherical joint. Results indicate that the spherical joint is extremely sensitive to eccentric loading and in most cases increases rather than decreases the bending stresses in the test, leading to lower apparent fracture strengths.

KEY WORDS: Kolsky Bar, Dynamic Fracture Strength, Dumbbell Specimens, 3D DIC, High-Speed Video

INTRODUCTION

The design of ceramic armor systems requires precise knowledge of the dynamic mechanical behavior of the ceramic armor materials, including dynamic compressive fracture strength, which is typically measured using a Split Hopkinson Pressure Bar, or Kolsky Bar (1). The fracture strength of brittle materials is very difficult to measure accurately because of the extreme sensitivity of the test results to mechanical alignment, and to end and surface effects. Researchers at the U.S. Army Research, Development and Engineering Command (RDECOM) have developed a dumbbell-shaped specimen to avoid premature fracture from end splitting due to tensile stresses developed during elastic punching, lateral strain mismatches, and/or friction that can occur when simple cylindrical specimens are used. As part of this research effort, the US Army RDECOM organized a limited round robin test effort involving three laboratories. In this paper we report on the testing performed at National Institute of Standards and Technology (NIST) as part of the limited round robin.

Because dumbbell specimens often have much larger aspect ratios compared to usually short cylindrical or cubic specimens used for dynamic testing, they might be susceptible to premature failure from bending. Alignment of the Kolsky bar interface is therefore crucial if bending is to be minimized during dynamic compression testing of dumbbell specimens. Alignment quality is usually checked by doing Kolsky bar tests with no sample and with the bar ends pressed tightly together. Perfect alignment results in complete transmission of the elastic loading pulse and no reflected pulse. While this condition is achievable in practice, in the present study we investigated whether bending stresses caused by small misalignment of the interface could allow large axial loads to be transmitted while minimizing bending in a dynamic compression test of a ceramic dumbbell specimen.

The present study compares fracture strength experiments performed with and without a spherical joint. The joint consists of a precisely-machined spherical interface placed on the transmission side of the specimen in Kolsky bar test. High-speed, three-dimensional digital image correlation (3D DIC) is used to measure the motion of the specimen and the platens during testing. From these measurements, specimen rotation is measured by comparing displacement vectors of the cylindrical ends of the dumbbell specimens with the original specimen axis. If the specimen rotates during the test, bending stresses are likely. The rotation observed in the DIC data are correlated to the apparent fracture strengths. Finally, the effect of the spherical joint on both rotation and apparent fracture stress is determined by comparing the results against similar experiments without the spherical joint.

Rhorer, Richard; Quinn, George; Mates, Steven P. ”Dynamic Compression Tests of Alumina Dumbbells Using a Spherical Joint.” Paper presented at Society for Experimental Mechanics 2019 Annual Meeting, Reno, NV, US. June 03, 2019 - June 06, 2019.

EXPERIMENTAL

Specimens provided by the US Army RDECOM were manufactured according to the specifications shown in Fig. 1. The test material was Coorstek[1] CAP3 alumina with a nominal Young's modulus of 370 GPa and a nominal compressive fracture strength of 5 GPa.

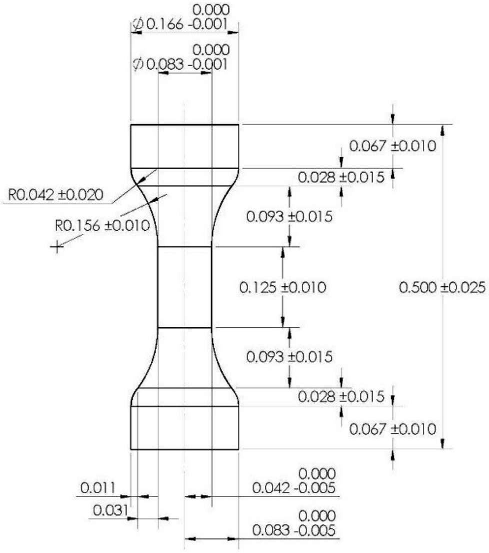


Fig. 1. Fabrication drawing for alumina dumbbell compression specimens. Dimensions in inches.

The NIST compression Kolsky Bar is pictured in Fig. 2 along with the results of a bar-to-bar test showing a small first reflection, indicative of imperfect interface alignment. Dimensions of the NIST Kolsky bar are given in Table 1 along with bar material properties and the longitudinal elastic wave speed in the bars. The Kolsky bar test was designed to fracture the samples in a single strike with a constant strain rate. For test design purposes, the breaking stress of this material was assumed to be 5 GPa, such that the fracture load of the specimens is estimated to be 17.8 kN. The fracture strain is estimated to be 0.0135 based on this fracture load and the nominal Young's modulus of 370 GPa. To break the sample, the striker length and impact velocity is selected to exceed both the peak load and the fracture strain. Based on a limited number of trials, a 203 mm long maraging steel striker bar (15 mm diameter) with an impact velocity of 9.9 m/s ± 0.1 m/s (± here is 2 standard deviations) were found to produce an adequate loading pulse. To achieve a constant strain rate, a ramped input pulse is created, following the recommendation in (1), with an annealed copper pulse shaper measuring 4.64 mm in diameter and 1.59 mm thick. Kolsky bar strain pulses are measured using dual 1000 Ω metal foil strain gages, one pair on each bar, located approximately midway along each bar. The strain gage bridge circuits are powered by a 24 V battery and the output is recorded at 2 MHz with a (14-bit) storage oscilloscope.

To protect the maraging steel bar ends from being damaged by ceramic fracture debris, square-shaped tungsten-carbide (WC) tool inserts (12 mm square by 3.22 mm thick) were placed on either side of the specimen as sacrificial protective elements. The WC tool inserts provided flat, hard surfaces that were well matched to the bar diameter. In some tests, round WC platens with stainless steel confining rings (12 mm diameter by 3 mm thick, impedance matched to the bars), leftover from previous studies, were used instead.

[1] Certain commercial equipment, instruments, or materials are identified in this paper to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

The spherical joint was constructed of 15 mm OD maraging steel. The mating surfaces, which were well matched and polished, had a radius of curvature of 17 mm, and the thickness of each piece was approximately 3 mm. The joint was placed directly behind the WC platen on the transmission side of the sample. The setup involving the spherical joints and a combination of square and round WC plates is shown in Fig. 3 (a). Interfaces between the bars and the spherical joint and WC platens were lubricated with heavy grease and the assembly was kept aligned with a small compressive force provided by elastic bands fixed between the bar and the bearing holders using capstan-style connectors.

Table 1. NIST Kolsky bar dimensions and materials.

Element	Quantity	Unit
Incident and Transmission Bar Length	1.5	m
Incident, Transmission & Striker Bar Diameter	0.015	m
Striker Bar Length	0.203	m
Bar Material	Maraging Steel (Un-Hardened)	
Bar Modulus	170	GPa
Bar Wave Speed	4600	m/s

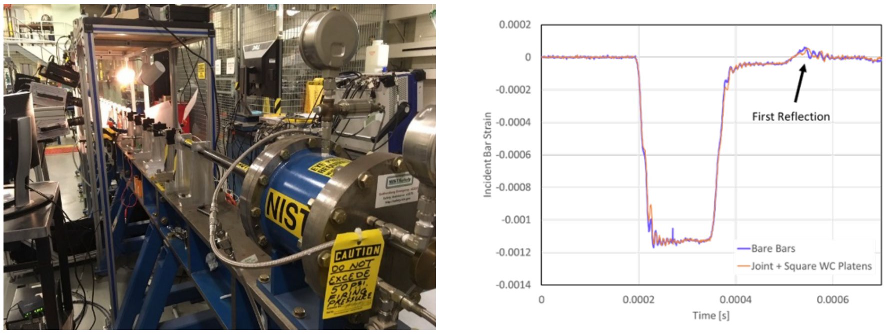
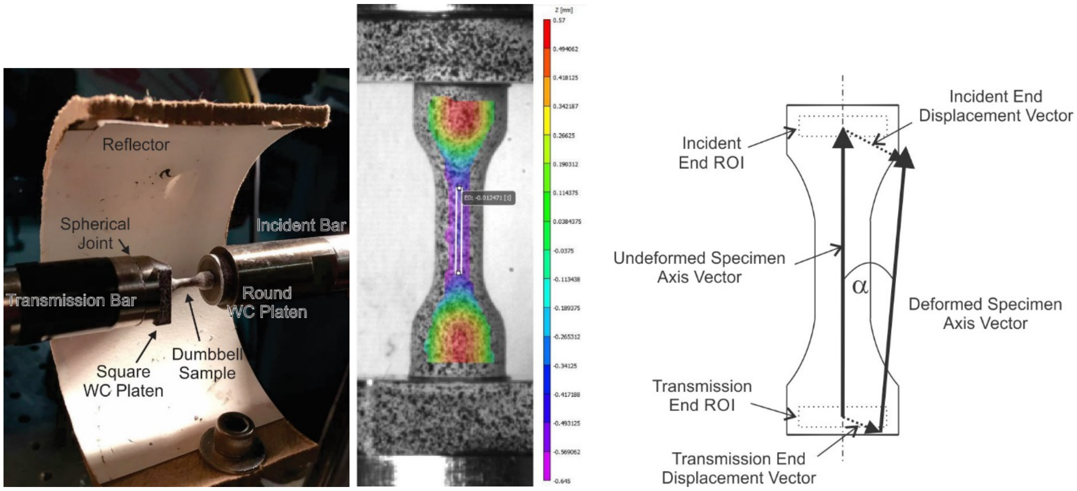


Fig. 2. Picture of the NIST compression Kolsky bar (left) and bar-to-bar test showing small first reflection indicative of imperfect interface alignment (right).

The imaging setup is pictured in Fig. **3** (a). Twin Photron SA1 high speed digital cameras with 90 mm macro lenses were used for 3D DIC measurements as well as to determine when and where the sample fails by videography. The lens apertures were set to the lowest setting (f/32) to maximize the image depth-of-field. The frame rate was 75 000 frames per second with a pixel resolution of 128 pixels by 352 pixels, and the exposure time was reduced to 1.76 µs using an electronic shutter on the cameras to eliminate motion blur prior to fracture. The geometric field of view was approximately 7.8 mm by 21.4 mm (about 16 pixels per mm), and the minimum object resolution, determined by the 1951 Air Force target method, was determined to be 0.0625 mm, indicating sharp focus. Illumination was provided by a single Photogenic PL2500 DR flash unit (871 W-s output over 0.001 s) positioned above the camera pair. A semi-circular diffuse reflector was added to improve illumination on the bottom side of the specimen and to provide a light background for the high-speed movies.

A sample DIC image is shown in Fig. 3 (b). Black speckles are produced on the while alumina sample using a fine, disperse spray of flat black spray paint. The included angle of the DIC cameras was 24 ° ± 2 °, and the working distance was 250 mm ± 25 mm. DIC measurements were obtained using Vic-3D from Correlated Solutions, with 15 pixel (0.94 mm) correlation windows (subsets) and a 3 pixel (0.19 mm) overlap. Correlations were determined using Gaussian subset weights with an optimized 6 tap interpolation scheme with normalized squared differences as the correlation criterion, according to the software settings. As this figure shows, the correlation region on the gage region is quite narrow, due to the high curvature of the specimen. Because of the limited number of DIC subsets available in the gage section, DIC measurements of sample strain in the gage section were obtained using the "virtual strain gage" method available in Vic-3D, which simply measures the relative displacement between two correlated subsets chosen along the gage section. The virtual gage is superimposed on the DIC image in Fig. 3 (b). The gage length in this example is 3.18 mm (51 pixels). The noise error for DIC displacement measurements, based on a set of 12 consecutive still images obtained under conditions identical to the real measurements, is a maximum of 1 µm and



(a) (b) (c)

Fig. 3. Sample, WC platens, spherical joint and reflector (a), DIC image of ceramic dumbbell showing correlation region and virtual gage used for strain measurement (b), and vector definitions for axis misorientation angle measurements (c).

is dominated by the out-of-plane displacement component. The noise is quite small compared to the typical maximum displacement measured prior to fracture (300 µm for the incident bell end, 160 µm for the trans bell end). Possible bias errors were not explored but are assumed to be small. The DIC parameters reported here are in partial fulfillment of the requirements laid out in the DIC Good Practice Guide (2).

RESULTS

Spherical Joint Tests with Ductile Metals

Experiments were conducted to ensure that the spherical joint and the WC protective platens did not introduce significant disturbance of the strain waves that might influence the data analysis in compression tests performed on actual samples. Previously, it was shown in bar-to-bar tests that the spherical joint and platens did not alter the wave transmission compared to tests without it (Fig. 2, right). Compression tests with and without the spherical joint and platens were conducted on samples of a brass alloy (26000) and on Ti-6Al-4V. Results of these tests are shown in Fig. 4. The results show that the presence of the additional elements does not change the stress-strain result for either material. We note that in these tests, the striker bar length and velocities were not identical, which caused the difference in total strains during the tests as evident in the figure. Regardless, the data agree quite well where they overlap in strain, indicating no significant disturbance of the waves through the joint and platens.

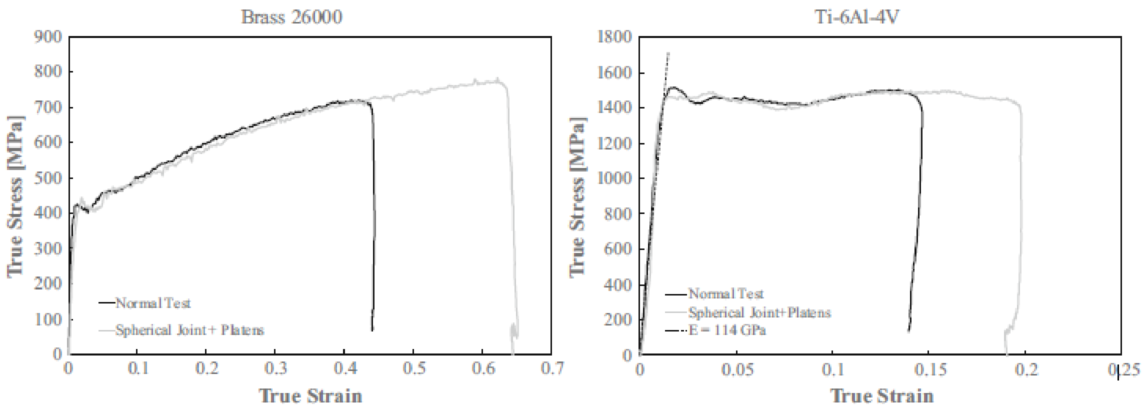


Fig. 4. Brass 26000 (left) and Ti-6Al-4V (right) measurements with and without spherical joint and WC platens. A nominal value of Young's Modulus (E) is plotted for Ti-6Al-4V for comparison with the data.

Alumina Dumbbell Compression Tests

The dynamic stress-strain response of the dumbbell samples is determined from the usual Kolsky bar strain wave analysis technique but with an adjustment to account for the strain concentration in the gage section of the dumbbell due to the reduced cross-sectional area there. Fig. 5 compares the strain-time data obtained from the adjusted wave analysis with DIC measurements directly on the gage section. Also shown in the figure is the platen strain, which is the overall dumbbell strain as determined from the measured displacements of the WC platens from DIC data. The DIC measurements indicated the gage strain is 1.4 times higher than the overall dumbbell strain due to the reduced cross-sectional area. In the case shown in Fig. 5, the adjusted wave strain is slightly higher than the adjusted platen strain, but the difference is within experimental uncertainty.

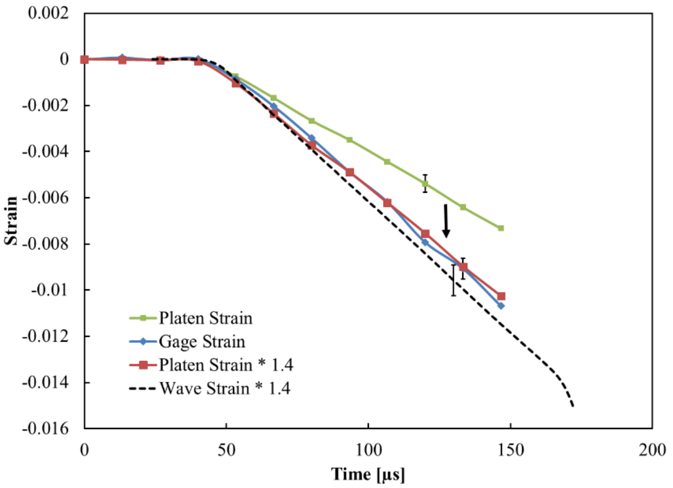


Fig. 5. Comparison of DIC Gage Strain (using the virtual gage method) with the overall strain of the dumbbell (Platen Strain from DIC-measured platen displacements), and the resulting correction factor of 1.4 applied to the Platen Strain data and the Wave Strain data. Error bars are determined using random error propagation and represent a coverage factor of 1.

Kolsky bar strain wave data showing the raw strain waves, equilibrium, strain-rate versus time and the stress-strain response for a dumbbell compression test that is considered valid (little obvious rotation observed from videography) are shown in Fig. 6. The apparent fracture stress, σ_{frac}, is determined from the peak load obtained from the Kolsky bar transmitted strain (ε_{frac}), and is determined from the following equation:

$$\sigma_{frac} = \frac{P_{frac}}{A_{gage}} = \frac{A_{bar}E_{bar}\varepsilon_{frac}}{\frac{\pi}{4}D_{gage}^{2}} \qquad (1)$$

P is load, E is Young's modulus, A is area. Subscript *gage* denotes dumbbell gage section properties, and *bar* denotes Kolsky bar properties. The uncertainty budget for the measurement of apparent fracture stress is given in Table 2, with the last entry being the calculated fractional uncertainty of the apparent fracture stress, considering independent, random error propagation. We note the apparent fracture stress uses the original cross-sectional area of the specimen, so it does not account for lateral expansion of the gage section due to Poisson effects.

Table 2. Uncertainty budget for fracture stress. Coverage factor = 1. δ denotes uncertainty.

Quantity	Uncertainty (Coverage Factor = 1)
$\delta A_{bar}/A_{bar}$	0.007
$\delta E_{bar}/E_{bar}$	0.012
$\delta\varepsilon_{frac}/\varepsilon_{frac}$	0.013
$\delta D_{gage}/D_{gage}$	0.004
$\delta\sigma_{frac}/\sigma_{frac}$	0.02

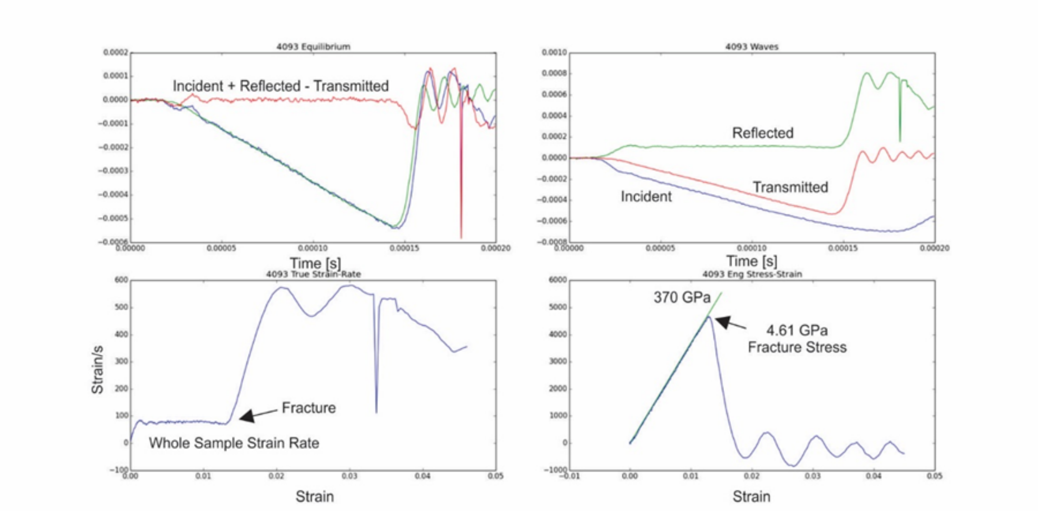


Fig. 6 Strain wave data from a dumbbell test that is considered valid. Upper left: comparison of incident, reflected and transmitted waves as a function of time. Upper right: raw incident, reflected and transmitted waves. Ordinate values of both plots in volts. Lower left: strain rate versus strain of the whole dumbbell sample, which is lower than the strain rate in the gage section by a factor of 1.4. Lower right: engineering stress-strain curve up to fracture (ordinate values in MPa).

Table 3 lists the results of the eight experiments where DIC data were available. Four tests used the spherical joint with WC platens and four tests used only WC platens. As the table shows, the apparent fracture stresses are generally lower when the spherical joint was employed, indicating that, rather than improving the test quality, the spherical joint actually makes things worse. Visual inspection of the high-speed movies revealed that the spherical joint appears to cause more significant rotation (bending) in the sample because of rotation in the joint. Fracture

Table 3. Test conditions and results[a].

Test ID Number	Ambient Temp [°C]	Ambient Relative Humidity (%)	Strain Rate [1/s] ± 5	Fracture Stress [MPa] ± 2 %	Fracture Strain ± 1.0 %	Spherical Joint?	α [Deg] ± 0.002
4085	23	18	115	3700	0.014	Y	0.130
4086	23	18	115	3530	0.013	Y	0.071
4087	23	18	122	3790	0.014	Y	0.260
4088	23	18	126	4460	0.013	N	0.050
4090	24	32	111	4040	0.013	Y	0.102
4091	24	32	112	4550	0.013	N	0.050
4092	24	32	109	4250	0.013	N	0.130
4093	24	32	113	4610	0.013	N	0.025

[a] Uncertainty estimates are determined with random error propagation techniques with a coverage factor = 1.

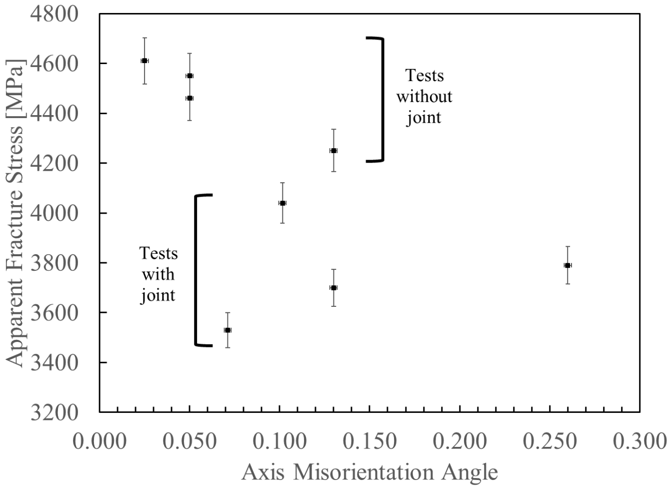


Fig. 7 Influence of measured axis misorientation angle just prior to fracture on apparent fracture stress in dynamic compression tests of brittle dumbbell specimens. Error bars reflect quantities given in Table 3.

strains, determined from the wave strain at fracture adjusted by the factor 1.4, are also listed in Table 3. A bias error in these strains is likely due to possible bending in the dumbbells during loading. Angles of rotation α are determined from DIC displacement data are also listed in Table 3, and their calculation is described below.

DIC measurements were used to quantify the deviation from uni-axial kinematics by measuring a misorientation angle, α, between the initial specimen axis and its axis just before fracture, as defined in Fig. 3 (c) and computed from the dot product of the two position vectors. Kinematically, in a uni-axial compression test of a dumbbell specimen, the axis of the dumbbell should remain parallel to the initial axis. If this is not the case, then the specimen is rotating and/or bending. Because the specimen is quite stiff, the amount of rotation and bending will be very small, even if bending stresses are high. As such, measuring rotation and/or bending requires very sensitive DIC displacement measurements. To examine whether rotation can be detected in our data, we compare the misorientation angle α with the apparent fracture stress, with the hypothesis being that the more the specimen rotates during the test, the larger the bending stress is likely to be, so that fracture strength should show a negative correlation with α (higher fracture stresses for lower misorientation angles). In addition, we expect, based on the visual assessment, that α will be larger for tests with the spherical joint compared to tests without the joint. Misorientation angles are reported in Table 3 just prior to fracture. The angles are quite small, as expected due to the high stiffness of the specimen and the Kolsky bar. While a random error analysis indicates that the uncertainty of the misorientation angle measurement is about two thousandths of a degree in the present case (based on the measured displacement noise components), bias errors have not been investigated, and a more thorough evaluation of the uncertainty magnitude is needed, for example using a very precise rotation stage. However, the results show larger misorientation angles when the spherical joint is used, confirming the earlier qualitative videography results. Further, as Fig. 7 shows, the apparent fracture stress increases as the misorientation angle decreases, which is also in-line with expectations. These results suggest that sensitive 3D DIC measurements may be able to provide useful information about the quality of a dynamic ceramic dumbbell compression test by measuring the degree of misorientation (bending) during the test. The simple misorientation analysis used here was motivated by the limited spatial and temporal resolution of the cameras on-hand; more advanced high-speed cameras with greater resolution at higher frame rates could provide other means to quantify bending, including a more accurate measure of the distortion of the dumbbell during loading and a more detailed measurement of the state of surface strain in the gage section. If bias sources can be more thoroughly quantified, a more definitive outlook on the usefulness of 3D DIC for improving the accuracy of dynamic compressive strength measurements of ceramic dumbbells can be made. The advantage of 3D DIC is rooted in the ability to measure out-of-plane displacements, which 2D DIC cannot provide. This capability could be used to guide the alignment of the Kolsky bar interface with dumbbell tests taken not quite to fracture, so allowing one to measure misorientation under load, saving valuable specimens until the bending is reduced to acceptable levels (possibly zero). The technique could also identify which fracture tests suffer from significant bending and should be discarded from the data set. Previously, the use of multiple gage-mounted strain gages has been suggested to monitor bending in dynamic compression tests on ceramics (1), although laboratories that already have access to very high-speed 3D DIC systems may consider using this optical method instead.

CONCLUSIONS

NIST's participation in a mini round robin involving dynamic fracture tests on dumbbell-shaped ceramic specimens has been described. A spherical joint was placed on the transmission side of the specimen to reduce bending loads on the specimen due to minor bar misalignment and thereby obtain more accurate fracture data. Instead, the joint resulted in lower apparent fracture strengths compared to tests performed without a joint. 3D DIC measurements of the specimen motion during testing showed a correlation between the amount of deviation from uni-axial kinematics and fracture strength, pointing to the possible usefulness of 3D DIC to monitor dynamic fracture test quality and improve overall test results by identifying invalid tests.

ACKNOWLEDGEMENTS

Dr. Jeffrey Swab, US Army RDECOM, for inviting us to participate in the round robin and for valuable discussions.

REFERENCES

1. Chen, W. and Song, B. *Split Hopkinson (Kolsky) Bar: Design, Testing and Applications.* New York : 2011, 2011. ISSN 0941-5122.
2. Jones, E.M.C. and Iadicola, M.A. *A Good Practices Guide for Digital Image Correlation.* s.l. : International Digital Image Correlation Society, 2018.

Reducing Effects of LO Cable Movement in Antenna and Long Distance VNA Measurements

David R. Novotny
Communications Technology Laboratory
National Institute of Standards and Technology
Boulder, Colorado, United States of America
david.novotny@nist.gov

***Abstract*— We evaluate a method for estimating and removing local oscillator cable drift in transmission measurements using a network analyzer. The geometric mean of the measured forward and reverse transmission, including drift, can be used to estimate the actual transmission without drift. This requires the measurement of passive, symmetrical transmission measured with a bi-directional two-port remote-mixing down-converting measurement system. This method is being used in antenna measurements where cable movement is unavoidable. It is viable for other calibrations and measurements using remote-mixing systems and frequency extenders at higher frequencies.**

Keywords*— *antenna measurement; calibration; drift; local oscillator; network analyzer

I. Introduction

A method has been shown to effectively remove the phase effects of local oscillator (LO) cable drift from remote-mixing vector network analyzer (VNA) measurements [1]. We present the circuit basis for extending this analysis for correcting complex, amplitude and phase, LO drift in measurements [2].

Characterizations of antennas (gain-extrapolation, polarization, and patterns) require antenna movement. Cable drift and movement are direct contributors to errors and increased measurement uncertainties [1-3]. These adverse effects become more significant with increasing frequency and electrical cable length, so frequency conversion is often used to minimize the effect of loss, and cable stress [1,4,5].

Transmitting the signal at the operating radio frequency (RF) though long cables can have large dynamic-range-limiting losses and induce large phase errors which may be both temperature and movement dependent [6]. A reference at lower LO frequency is used to down-convert the higher test RF to a lower intermediate frequency (IF). This lower frequency LO reference and IF measurement can travel farther with better fidelity.

The downside of this approach is that using a remote LO creates a non-linear dependence between LO drift and the desired signal measurement. Ideally, keeping the LO drift small minimizes these errors. Complex systems have been built to measure and compensate for RF, LO and IF drift, which have dramatically reduced these effects [4]. Rather than treating LO drift as a random uncertainty, we estimate it through measurements [1,3,7]. We present a simple method using the geometric mean of forward and reverse measurements through a reciprocal system to estimate actual transmission.

II. Typical Antenna Measurements

Single direction measurement systems (i.e. only S_{21} or S_{12}) are typically used for antenna measurements as cost, speed and complexity are driving factors. Mixers and cabling for both directions can often be twice the cost, can add more hardware near the antenna (possibly increasing RF scattering) and may require more complex calibration to reduce systematic losses of the measurement system.

Fig. 1. shows several conventional RF setups used in antenna measurements. Fig. 1(a) depicts a common far-field range. Here differential changes in the LO feed network show up directly as differences in the measured IF signal. Fig. 1(b) shows no external frequency conversion. This can lead to excessive loss and RF amplitude phase changes due to the cabling, which become worse as frequency and/or range of movement increases. Fig. 1(c) depicts a single side down conversion. Fig. 1(d) shows a source and receiver conversion RF setup. The LO is split and may drift differently for the reference and antenna under test (AUT).

III. Dual-Direction Remote-Mixing

The National Institute of Standards and Technology (NIST) is using robotic systems to move antennas during the measurement of radiation characteristics from 1-500 GHz. To minimize cable transmission losses at higher frequencies and allow physical separation between the VNA and the antennas, we employ a remote-mixing, frequency-conversion scheme available from many manufacturers. Fig. 2(c) shows the signal

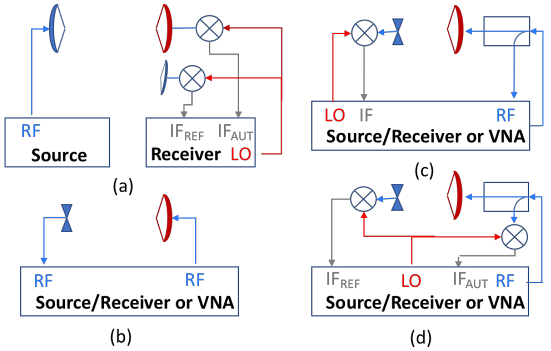


Figure 1. Typical single direction test setups used in antenna measurements. RF signal paths are shown in blue, LO reference paths are red, and IF signal paths are shown in gray.

Novotny, David. "Reducing Effects of LO Cable Movement in Antenna and Long Distance VNA Measurements." Paper presented at 2019 Joint International Symposium on Electromagnetic Compatibility, Sapporo and Asia-Pacific International Symposium on Electromagnetic Compatibility (EMC Sapporo/APEMC), Sapporo, JP. June 03, 2019 - June 07, 2019.

block diagram of our measurement system. To determine the desired RF signal quantities, a_1, b_1, a_2, b_2, from the ideal VNA inputs, a_1', b_1', a_2', b_2', the error network, Fig. 2(a), needs to be characterized by a calibration. Large physical scans (>100λ) at these frequencies can stress cables that result in changes in receiver measurements not due to the actual antenna measurement. We will show that this is due largely to a mixer network which responds non-linearly to LO cable movement, Fig 2(b).

A. Assumptions, Limitations and Simplications

The major assumptions and limitations for the initial analysis are:

- Both directions are being measured and the antenna measurement is reciprocal, i.e. $S_{21}=S_{12}$. This limits the analysis for active antennas and usage of isolators/circulators.
- LO cable change is the dominant form of drift, i.e. we ignore thermal effects of the VNA or extender head.
- The IF cables are operating at a low enough frequency that amplitude and phase changes due to cable movement at IF can be ignored.
- The RF signal drift is removed by the ratioed S-parameter analysis.
- The system does not systematically change during the forward and reverse scattering parameter measurements. In practice, this is accomplished by limiting the movement of the antennas to no more than $\lambda_{RF}/50$ during the RF measurement.
- There is no movement/drift between the calibration plane and the antenna. This implies avoiding flexible connections between the frequency converter and the antenna. This drift effect can often be more significant than LO cable drift [3].

B. Signals to the Down-Converting Mixers

The standard error model for VNA calibration is dependent on a stable signal path between the test port and the VNA receivers. We assume the error networks in Fig. 2(c) are stable and the mixer networks transmit the coupler output signals, a_1', b_1', a_2', b_2', to the IF inputs of the VNA a_1'', b_1'', a_2'', b_2''.

We will propagate real-time signals from the coupler outputs through the mixer network to determine the effective loss and delay. The error network output signal, a_1', at the RF port of the mixer can be written as:

$$V_{A6}(t) = V_{RFa1}(t) = A_{a1}\sin(\omega_{RF}t + \phi_{a1}) \qquad (1)$$

Where ω_{RF} is the angular frequency of the RF test signal, t is time, and $A_{xi}$and ϕ_{xi} are the amplitude and phase of the signals impingent on the mixers for port i and wave direction x.

The LO signal at the port 1 frequency converter, A2, comes from source, A1, where it may be split and transmits through the LO cable to A2. The received LO signal at the frequency converter,$V_{A2}(t)$, is then sent to a limiting amplifier (depicted in Fig. 2(c) as a saturated amplifier and limiter) and typically filtered to reduce amplifier induced harmonics.

$$V_{A3}(t) = E_{lim1}(L_{LC1}A_{LO})\sin(\omega_{LO}t + \phi_{LO1} + \phi_{LC1} + \phi_{lim1}) \qquad (2)$$

where$L_{LC1}$and ϕ_{LC1}represent the LO cable loss and delay and $E_{lim1}(x)$ and ϕ_{lim1} are the amplitude transfer function and phase delay between the input to the frequency extender and the input to the *xn* LO up-converting mixer. The output of the up-converting mixer is split and sent to the down-converting mixers at A4, repented by:

$$V_{A4}(t) = C_{MUa1}E_{lim1}(L_{LC1}A_{LO}) \\ \sin[n(\omega_{LO}t + \phi_{LO1} + \phi_{LC1} + \phi_{lim1}) + \phi_{MUa1}] \qquad (3)$$

where C_{MUa1} and ϕ_{MUa1} represent the loss and phase delay, including the *xn* mixer's conversion and splitting loss, between the *xn* mixer and the input to the down-converting mixers. The *xn* mixer not only upconverts the LO frequency by a factor of *n*, but also multiplies any phase shifts after the mixer by the same factor.

The output of the down-converting mixer, A8, is the product of the two inputs to the mixer with a conversion loss:

$$V_{A8}(t) = \frac{C_{MDa1}A_{a1}C_{MUa1}E_{lim1}(L_{LC1}A_{LO})}{2} \qquad (4)$$
$$cos[(\omega_{RF} - n\omega_{LO})t - n(\phi_{LO1} + \phi_{LC1} + \phi_{lim1}) + \phi_{MUa1} \\ + \phi_{a1} + \phi_{MDa1}]$$

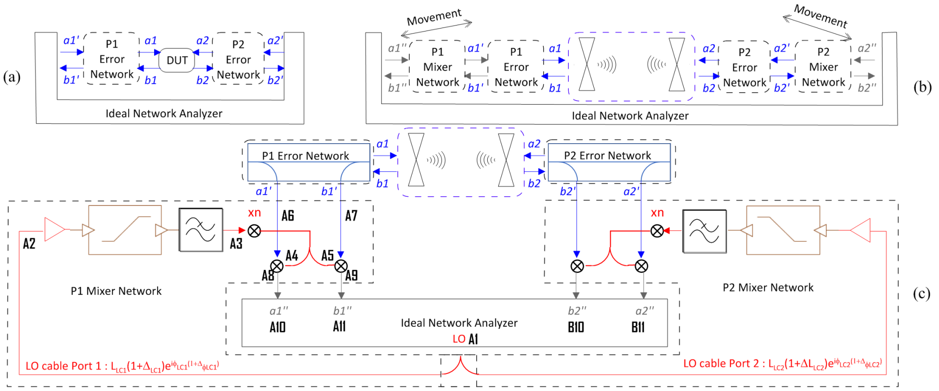


Figure 2. Block diagrams of the ideal VNA calibration with a static error network between the VNA and the test ports(a), (b) a changing error network due to movement such as during an antenna test, and (c) a signal flow model of the down-conversion to get the signal between the frequency extenders and the VNA.

Novotny, David. "Reducing Effects of LO Cable Movement in Antenna and Long Distance VNA Measurements." Paper presented at 2019 Joint International Symposium on Electromagnetic Compatibility, Sapporo and Asia-Pacific International Symposium on Electromagnetic Compatibility (EMC Sapporo/APEMC), Sapporo, JP. June 03, 2019 - June 07, 2019.

where C_{MDa1} and ϕ_{MDa1} are the mixer down-conversion loss and phase delay between A6 and A8. The upper mixing product, $(\omega_{RF} + n\omega_{LO})$, is filtered out by the mixer and the IF cabling leaving the IF signal $(\omega_{RF} - n\omega_{LO})$. The signal at A10 or the VNA receiver, a_1'', is the signal at A8 with the IF cable loss, L_{IFa1}, and phase delay, ϕ_{IFa1}, at the IF frequency.

$$V_{A10}(t) = a_1'' = \frac{L_{IFa1}C_{MDa1}A_{a1}C_{MUa1}E_{lim1}(L_{LC1}A_{LO})}{2} \\ cos[(\omega_{RF} - n\omega_{LO})t - n\phi_{LC1} + \phi_{a1} + \phi_{MUa1} + \phi_{MDa1} \\ + \phi_{IFa1}] \tag{5}$$

If we can assume thermal and mechanical stability except for the LO cable, and that at IF, the movement of the IF cables cause minimal change in complex loss, we can simply (5) to:

$$V_{A10}(t) = a_1'' = A_{a1}K_{a1}E_{lim1}(L_{LC1}A_{LO})\, cos[(\omega_{RF} - \\ n\omega_{LO})t - n\phi_{LC1} + \phi_{a1} + \phi_{Ka1}] \tag{6}$$

IV. How LO Changes Affect a Calibration

VNA calibration is accomplished by employing standards with “known” scattering parameters and measuring the response of the VNA with the attached error network and then removing the effects of the error network [8-10]. (6) shows a linear amplitude and phase relationship between the RF signal input and the VNA receivers as long as the LO cable stays constant. However, we see that LO cable changes in L_{LC1} and ϕ_{LC1} affect the relationship between a_1'' and a_1'.

A. Response Calibration

The simplest two-port error network characterization is using the response or “thru” calibration. The measurement ports are connected via a known transmission network, often a flush zero-loss device, and the source and transmission are measured relative to the“thru”.

To minimize notation, we can re-write (6) at the IF in amplitude and phase notation with the frequency suppressed:

$$\begin{aligned} a_1'' &= a_1'K_{a1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1} + \phi_{Ka1}) \\ b_1'' &= b_1'K_{b1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1} + \phi_{Kb1}) \\ a_2'' &= a_2'K_{a2}E_{lim2}(L_{LC2}A_{LO})\angle(-n\phi_{LC2} + \phi_{Ka2}) \\ b_2'' &= b_2'K_{b2}E_{lim2}(L_{LC2}A_{LO})\angle(-n\phi_{LC2} + \phi_{Kb2}) \quad . \end{aligned} \tag{7}$$

The reference value of the response calibration, with transmission $T\angle\phi_T$, is calculated at what is assumed to be the nominal state of the LO cables.

$$S_{21_{ref}} = \frac{b_2}{a_{1_{ref}}} = T\angle\phi_T = \frac{b_2''}{a_1''}_{ref} K_{21_{ref}} = \\ K_{21_{ref}} \frac{b_2'}{a_{1_{ref}}'} \frac{K_{b2}E_{lim2}(L_{LC2}A_{LO})\angle(-n\phi_{LC2}+\phi_{Kb2})}{K_{a1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1}+\phi_{Ka1})} \tag{8}$$

$$S_{12_{ref}} = \frac{b_2}{a_{1_{ref}}} = T\angle\phi_T = \frac{b_1''}{a_2''}_{ref} K_{12_{ref}} = \\ K_{12_{ref}} \frac{b_1'}{a_{2_{ref}}'} \frac{K_{b1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1}+\phi_{Kb1})}{K_{a2}E_{lim2}(L_{LC2}A_{LO})\angle(-n\phi_{LC2}+\phi_{Ka2})}$$

A calibrated response measurement can be made by solving for $K_{21_{ref}}$

$$K_{21_{ref}} = T\angle\phi_T / \left(\frac{b_2''}{a_1''}_{ref}\right), K_{12_{ref}} = T\angle\phi_T / \left(\frac{b_1''}{a_2''}_{ref}\right) \tag{9}$$

If a differential change is now added to the LO cables, as happens in moving antenna measurements, we see an amplitude and phase change (e.g. the port 1 LO cable transmission, $L_{LC1}\angle\phi_{LC1}$, becomes $L_{LC1}(1+\Delta_{LC1})\angle\phi_{LC1}(1+\Delta\phi_{LC1})$). This effects the single-direction measurement of S_{21}:

$$S_{21_{meas}} = \frac{b_2}{a_1} \sim \frac{b_2''}{a_1''} K_{21_{ref}} = \frac{b_2''}{a_1''} T\angle\phi_T / \left(\frac{b_2''}{a_1''}_{ref}\right) = \frac{b_2'}{a_1'} \\ \frac{K_{b2}E_{lim2}(L_{LC2}(1+\Delta_{LC2})A_{LO})\angle(-n\phi_{LC2}(1+\Delta\phi_{LC2})+\phi_{Kb2})}{K_{a1}E_{lim1}(L_{LC1}(1+\Delta_{LC1})A_{LO})\angle(-n\phi_{LC1}(1+\Delta\phi_{LC1})+\phi_{Ka1})} \\ T\angle\phi_T \frac{a_1'}{b_{2_{ref}}'} \left[\frac{K_{a1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1}+\phi_{Ka1})}{K_{b2}E_{lim2}(L_{LC2}A_{LO})\angle(-n\phi_{LC2}+\phi_{Kb2})}\right]. \tag{10}$$

If there are no changes in the LO cables, (10) reduces to the standard calibrated thru result. However, changes in either of the LO cables will directly translate into changes in the measured transmission: the standard side-effect of using a single-direction measurement. So, the measurement of just S_{21} or S_{12} using the VNA receivers may be compromised.

However, if we can assume that the antenna measurement is reciprocal, $S_{21} = S_{12}$, or $b_2/a_1 = b_1/a_2$, we can look at (10) and see that S_{12} has an inverse relationship with LO changes from S_{21}. If both directions are measured and the geometric mean of the results are taken, it can cancel out the LO drift effects:

$$\sqrt{S_{21_{meas}}S_{12_{meas}}} = \sqrt{\frac{b_2''}{a_1''}K_{21_{ref}} \frac{b_1''}{a_2''}K_{12_{ref}}} \\ = \sqrt{\frac{b_2''}{a_1''}\left(T/\left(\frac{b_2''}{a_1''}_{ref}\right)\right)\frac{b_1''}{a_2''}\left(T/\left(\frac{b_1''}{a_2''}_{ref}\right)\right)} = \left\{\frac{b_2'}{a_1'}\right. \\ \left[\frac{K_{b2}E_{lim2}(L_{LC2}(1+\Delta_{LC2})A_{LO})\angle(-n\phi_{LC2}(1+\Delta\phi_{LC2})+\phi_{Kb2})}{K_{a1}E_{lim1}(L_{LC1}(1+\Delta_{LC1})A_{LO})\angle(-n\phi_{LC1}(1+\Delta\phi_{LC1})+\phi_{Ka1})}\right] \\ T\angle\phi_T \frac{a_1'}{b_{2_{ref}}'} \left[\frac{K_{a1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1}+\phi_{Ka1})}{K_{b2}E_{lim2}(L_{LC2}A_{LO})\angle(-n\phi_{LC2}+\phi_{Kb2})}\right] \frac{b_1'}{a_2'}` \\ \left[\frac{K_{b1}E_{lim1}(L_{LC1}(1+\Delta_{LC1})A_{LO})\angle(-n\phi_{LC1}(1+\Delta\phi_{LC1})+\phi_{Kb1})}{K_{a2}E_{lim2}(L_{LC2}(1+\Delta_{LC2})A_{LO})\angle(-n\phi_{LC2}(1+\Delta\phi_{LC2})+\phi_{Ka2})}\right] \\ \left. T\angle\phi_T \frac{a_2'}{b_{1_{ref}}'} \frac{K_{a2}E_{lim1}(L_{LC2}A_{LO})\angle(-n\phi_{LC2}+\phi_{Ka2})}{K_{b1}E_{lim1}(L_{LC1}A_{LO})\angle(-n\phi_{LC1}+\phi_{Kb1})}\right\}^{\frac{1}{2}} \tag{11}$$

$$= \sqrt{\frac{b_2'}{a_1'}\left(T/\left(\frac{b_2'}{a_{1_{ref}}'}\right)\right)\frac{b_1'}{a_2'}\left(T/\left(\frac{b_1'}{a_{2_{ref}}'}\right)\right)} = \sqrt{\frac{b_2}{a_1}\frac{b_1}{a_2}} = \sqrt{S_{21}S_{12}} = S_{12} = S_{21}$$

B. Practial Implementation

Equation (11) shows that measured geometric mean of the forward and reverse measurements can be an estimate of the transmission through a passive, bi-directional, down-converting receiver setup. However, (11) has multiple solutions to the phase of the transmission signal [1]. If the net LO phase changes, $n(\Delta\phi_{LC1} - \Delta\phi_{LC2}) > \pi$, there can be an angle ambiguity, Fig. 3, because there are two solutions to square root in (11) that are separated by π. The correct branch can be found by keeping the movements small between RF measurements so to track the best branch, or by taking a broad frequency sweep so the correct solution can be chosen to keep the phase consistent.

Novotny, David. "Reducing Effects of LO Cable Movement in Antenna and Long Distance VNA Measurements." Paper presented at 2019 Joint International Symposium on Electromagnetic Compatibility, Sapporo and Asia-Pacific International Symposium on Electromagnetic Compatibility (EMC Sapporo/APEMC), Sapporo, JP. June 03, 2019 - June 07, 2019.

C. Reflection Measurments

Since reflection measurements use a single port, they only require one LO cable. Re-writing (11) for one-port results in:

$$S_{11} = \frac{b_1}{a_1} \propto \frac{b_1'}{a_1'} \frac{K_{a_1'} E_{limA}(L_{LOA} A_{LO}) \angle\left(-n\phi_{LOA} + \phi_{a_1'}\right)}{K_{b_1'} E_{limA}(L_{LOA} A_{LO}) \angle\left(-n\phi_{LOA} + \phi_{b_1'}\right)}. \quad (12)$$

(12) shows that the single-port scattering parameters are primarily dependent on one LO cable, so drift effects are normalized and reduced.

D. Other Calbration Methods

As the dominant term in the transmission error correction in most two-port calibrations is the transmission coefficient of the reference thru, a form of (10) is relevant in systems similar to Fig. 2(c). Estimating the transmission using $\sqrt{S_{21}S_{12}}$ with LO cable drift can be used for other two-port calibration methods such as TRL[9], SOLT, and “Unknown-Thru”[9]. A major advantage of using the “unknown-thru” in antenna measurements is that VNA calibrations can be done with the test antennas in place: this avoids the need for a mechanical-thru connection between the test ports, limiting port movement when measuring the standards during RF calibration.

V. Measured Data

We performed a two-port LRL calibration in WR-05 from 140-180 GHz. The *xn* up-conversion factor was 12 and the IF was set at 100 MHz. A power calibration was performed to ensure the LO was in the proper power range for the frequency converters. Fig 4. Shows the results viewing the flush-thru after calibration, so all data should be 0 dB and 0°. Moving the LO cable for each port shows minimal amplitude change but large phase changes. Using the measured $\sqrt{S_{21}S_{12}}$ to estimate S_{12} shows promise even when a 2-cm 2-dB pad is inserted into the LO line.

VI. Uncertainity Implications

The random sampling noise in $\sqrt{S_{21}S_{12}}$ has the potential to be larger than in S_{12} or S_{21}. So, for systems with small levels of LO stress this may increase uncertainty in the transmission results.

VII. Conclusion

The geometric mean of the measured forward and reverse transmission, in a system that has drift due to LO cable movement, can be used to estimate the actual drift-free transmission. This method requires a measurement of a passive and reciprocal system using a bi-directional remote down-converting receiver system. This method is especially useful when LO cable stresses are expected to be electrically large, as in antenna measurements and high-frequency measurements where movement of the test ports is much greater than a wavelength.

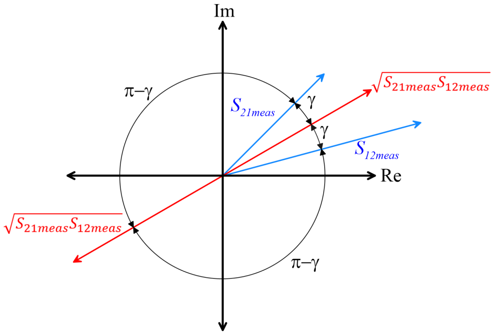


Figure 3. Geometric mean of the forward and reverse transmission measurement showing both solutions.

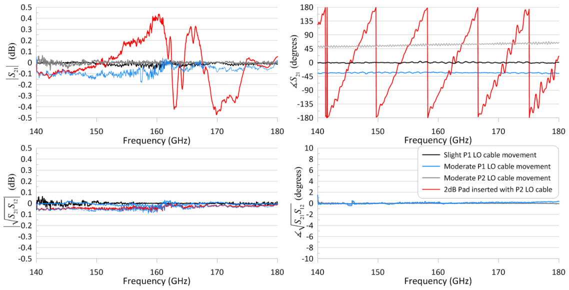


Figure 4. Transmission estimation of a flush thru using the measured $S_{21}$and $\sqrt{S_{21}S_{12}}$ with LO cable movement and inserting a 2 dB pad into the LO line. The majority of amplitude errors are reduced by the limiter in the converter, phase and residual amplitude errors are reduced by $\sqrt{S_{21}S_{12}}$.

References

[1] Arsenovic, A. (2017). A Method to Remove the Effects of LO Drift from Vector Network Analyzer Measurements. 10.13140/RG.2.2.23380.60806.

[2] D. R. Novotny, A. J. Yuffa, R. C. Wittmann, M. H. Francis, J. A. Gordon, “Some Advantages of Using Bi-directional S-Parameters in Near-Field Measurements,” in 40th Proc. Antenna Meas. Techn. Assoc. Conf., Nov. 2018.

[3] A. Lewandowski and D. Williams, "Characterization and modeling of random vector network analyzer measurement errors," MIKON 2008 - *17th International Conference on Microwaves, Radar and Wireless Communications*, Wroclaw, 2008, pp. 1-4.

[4] S. L. Dvorak and B. K. Sternberg, "Removal of time-varying errors in network-analyser measurements: signal normalisation and test results," *IEE Proceedings - Science, Measurement and Technology*, vol. 149, no. 1, pp. 31-36, Jan. 2002.

[5] I. Ida, H. Yoshimura and K. Ito, "Reduction of drift error of a network analyser in small antenna measurement," in IEE Proceedings - Microwaves, Antennas and Propagation, vol. 148, no. 3, pp. 188-192, June 2001. doi: 10.1049/ip-map:20010404.

[6] J. Kang, J. Kim, N. Kang and D. Kim, "Antenna measurement using S-parameters," *2012 Conference on Precision electromagnetic Measurements*, Washington, DC, 2012, pp. 658-659. doi: 10.1109/CPEM.2012.6251101

[7] Keysight, “Measurement Uncertainty of VNA Based TDR/TDT Measurement, Apllicaiton Note 5990-8406EN” USA, Aug 2014.

[8] S. Rehnmark, "On the Calibration Process of Automatic Network Analyzer Systems (Short Papers)," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 22, no. 4, pp. 457-458, Apr. 1974. doi: 10.1109/TMTT.1974.1128250

[9] Andrea Ferrero and Umberto Pisani, “Two-Port Network Analyzer Calibration Using an Unknown ‘Thru’,” IEEE Microwave and Guided Wave Letters, Vol. 2, No. 12, December 1992, pp 505-506.

[10] G. F. Engen and C. A. Hoer, "Thru-Reflect-Line: An Improved Technique for Calibrating the Dual Six-Port Automatic Network Analyzer," in *IEEE Transactions on Microwave Theory and Techniques*, vol. 27, no. 12, pp. 987-993, Dec. 1979. doi: 10.1109/TMTT.1979.1129778

Novotny, David. "Reducing Effects of LO Cable Movement in Antenna and Long Distance VNA Measurements." Paper presented at 2019 Joint International Symposium on Electromagnetic Compatibility, Sapporo and Asia-Pacific International Symposium on Electromagnetic Compatibility (EMC Sapporo/APEMC), Sapporo, JP. June 03, 2019 - June 07, 2019.

Implementing a Protocol Native Managed Cryptocurrency

Peter Mell
National Institute of Standards and Technology
Gaithersburg MD, USA
peter.mell@nist.gov

Aurelien Delaitre
Prometheus Computing
New Market MD, USA
aurelien.delaitre@nist.gov

Frederic de Vaulx
Prometheus Computing
New Market MD, USA
frederic.devaulx@nist.gov

Philippe Dessauw
Prometheus Computing
New Market MD, USA
philippe.dessauw@nist.gov

***Abstract*—Previous work presented a theoretical model based on the implicit Bitcoin specification for how an entity might issue a protocol native cryptocurrency that mimics features of fiat currencies. Protocol native means that it is built into the blockchain platform itself and is not simply a token running on another platform. Novel to this work were mechanisms by which the issuing entity could manage the cryptocurrency but where their power was limited and transparency was enforced by the cryptocurrency being implemented using a publicly mined blockchain. In this work we demonstrate the feasibility of this theoretical model by implementing such a managed cryptocurrency architecture through forking the Bitcoin code base. We discovered that the theoretical model contains several vulnerabilities and security issues that needed to be mitigated. It also contains architectural features that presented significant implementation challenges; some aspects of the proposed changes to the Bitcoin specification were not practical or even workable. In this work we describe how we mitigated the security vulnerabilities and overcame the architectural hurdles to build a working prototype.**

***Index Terms*—Fiat Currency, Cryptocurrency, Bitcoin**

I. Introduction

The United States National Institute of Standards and Technology developed an architecture for a managed cryptocurrency that has many of the features of electronic fiat currencies and includes a governing entity [1]. It is intended to combine the strengths of both fiat currencies and cryptocurrencies. In doing this, it deviates from the goals of most cryptocurrencies by introducing concepts such as central banking, law enforcement, and identity proofed accounts. It also deviates from a government controlled fiat currency world in denying the currency administrator absolute power over financial controls. It enables a currency administrator to enact policy to create a specific cryptocurrency instance from the architecture, usually with immutable configurations that even the administrator cannot violate. This can promote public trust in the currency since the limits to the administrator's power are immutably recorded on the associated blockchain. The architecture uses a public permissionless blockchain approach whereby the administrator's actions are completely transparent. Furthermore, a public set of miners maintaining the blockchain can prevent the administrator from performing unauthorized actions. At the same time, the cryptocurrency is designed to prevent the public miners from taking control from the administrator or from preventing the administrator's transactions from being processed. This architecture thus creates a 'balance of power' between the administrator and the public miners. Additional features include adding role attributes to cryptocurrency acounts that represent fiat currency entities (e.g., commercial banks, central banks, and law enforcement) such that there is created a tree based hierarchy of nodes with roles for all users of the cryptocurrency.

A major limitation to the approach is that it was presented only as a theoretical architecture. It demonstrated what might be possible to create through modest forks to existing cryptocurrencies, specifically using Bitcoin [2] [3] [4] as an example. The empirical work was limited to proposing changes to the implicit Bitcoin specifications in [5] and [6] to add the features necessary for this 'balance of power' managed cryptocurrency approach. No code was developed and no implementation was tested. The ability of [1] to modify the Bitcoin specification to add the needed features indicated that a managed cryptocurrency might be able to be built through a modest fork of an existing cryptocurrency, but it lacked a proof-of-concept prototype built as a protocol native implementation.

In this work, we set out to build such a prototype as an applied research endeavor. We tested whether or not such a managed cryptocurrency system could be built through modest modifications to the code base of an existing cryptocurrency. In this way we explored how to create a protocol native managed cryptocurrency built into the blockchain platform itself and explore the advantages of this approach. This was non-trivial as we did not simply create a token on top of another cryptocurrency. We also wanted to see if this could be done efficiently, with only a modest amount of programming effort (we scoped using half a person year, in part due to resource constraints). We chose to use Bitcoin since [1] described their theoretical model through proposing changes to Bitcoin. We wanted to discover the complexity of modifying Bitcoin to require identity proofing of accounts, establish accounts with roles, enable law enforcement functions, enable central banking functions, and create and visualize a hierarchy tree of accounts that specifies the scope of control of the various management and law enforcement nodes.

An unattributed quote says that 'theory is when you know everything but nothing works.' Yogi Berra said, 'in theory there is no difference between theory and practice. But, in practice, there is.' We found these statements to be true with regard to our implementation of the theoretical work. We discovered that the theoretical model contains several

Mell, Peter; Delaitre, Aurelien; de Vaulx, Frederic; Dessauw, Philippe. "Implementing a Protocol Native Managed Cryptocurrency." Paper presented at The Fourteenth International Conference on Software Engineering Advances (ICSEA 2019), Valencia, GA, ES. November 24, 2019 - November 28, 2019.

vulnerabilities and security issues that needed to be mitigated. It also contains architectural features that presented significant implementation challenges; some aspects of the proposed changes to the Bitcoin specification were not practical or even workable. We thus had to augment the material in [1] in order to achieve a functional and secure system, especially in areas such as preserving the balance of power, law enforcement powers, management node powers, bootstrapping the system, and the needed movement of accounts within the node hierarchy (e.g., when an account holder changes their account manager). We also encountered difficulties using the Bitcoin code base which necessitated design changes not foreseen in [1]. However, in the end we discovered that it was possible to modestly modify Bitcoin to implement this 'balance of power' managed cryptocurrency approach and to do it with a relatively low amount of programming effort.

In summary, we showed that the theoretical architecture provided by [1] works and can be implemented efficiently. However, we had to change, refine, and augment the original design in order to make it function. This paper describes these changes and the final prototype implementation which we have made publicly available on GitHub (any mention of commercial products is for information only; it does not imply recommendation or endorsement). Note that due to resource constraints, our prototype is not a full implementation. The largest limitation is that the cryptocurrency policy configuration is static, while the full design in [1] permits dynamic policy changes. While not all features were implemented, the core functionality was enabled to provide confidence that the system could be efficiently constructed.

The rest of this paper is organized as follows. Section II presents the theoretical architecture from [1] and discusses relevant Bitcoin architectural features. Section III discusses the vulnerabilities and security issues we discovered in the architecture. Section IV discusses the architectural hurdles that we had to overcome. Section V outlines how we created our prototype system and Section VI presents the related work. Section VII discusses our future plans for the system and Section VIII concludes.

II. Theoretical Architecture

The research in [1] provides an architecture that can be instantiated into a cryptocurrency instance through specifying a specific policy configuration. The policy parameters enable or disable feature sets while specifying parameters for cryptocurrency operation. The financially related parameters are just examples of what could be (e.g., limits on money production) and are not intended to be exhaustive given that the identification of financial controls is a related but separate research area. In this architecture, anyone can create an account, but an account cannot do anything unless it is granted one or more roles. The initial block on the blockchain has a 'genesis transaction' that grants roles to the root administrator account and all future role assignments spring from this initial root account. The root account grants roles to other accounts, and those accounts in turn may grant roles to accounts. This sets up a hierarchy of accounts in a tree structure with the root account (or node) being the most authoritative.

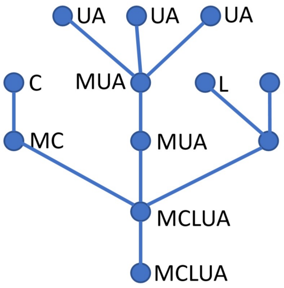


Fig. 1. Example Managed Cryptocurrency Hierarchy (from [1])

The initial root node is given all possible roles so that it can propagate these roles to other accounts. Of particular import is the 'M' currency manager role that enables an account to give its roles to other accounts (or withdraw granted roles) and to modify cryptocurrency policy. Other roles include 'U' user, 'A' account manager, 'C' central banker, and 'L' law enforcement. Their abilities are summarized in [1] as follows:

- 'The U role enables an account to receive and spend coins. An account for which the U role has been removed has its funds frozen.
- The A role enables a node to create accounts with the U role (and only the U role). It may also remove the U label for its descendants.
- The C role enables the creation of new coins (apart from the block mining rewards).
- The L role enables an account to forcibly move funds between accounts, to remove the U label, and to restore a previously removed U label. However, these actions can only be performed against nodes with the same or greater distance from the root.'

Note that in this model the currency administrator controls the root manager node and thus controls the privileges of all other nodes participating in the system. It can thus ensure that the A nodes perform identity proofing of U nodes (if desired). This can enable law enforcement, at least with a court order, to identify individuals within the system. This goes counter to the trend in cryptocurrencies where privacy and non-traceability are key objectives. An example node hierarchy tree with role assignments is shown in Figure 1.

There are three types of transactions that enable accounts with roles to perform their functions: coin transfer mode, role change mode, and policy change mode. A large portion of [1] specifies how to modify the nValue field in Bitcoin (which normally specifies the amount of coin to transfer) to enable the role and policy change functionality while still enabling coin transfer (but now only between accounts with the U role).

Lastly, there are two possible security models. There is an independent mining model where the miners are truly independent from the currency administrator, but they could then as a group deny the inclusion of management transactions (i.e., role changes and policy changes). This would be similar to a 51 % attack [7] being launched against Bitcoin. To prevent this there is also a dependent mining model where the miners must include a certain number of management transactions every so many blocks. This can prevent a large group of miners from being able to revolt and exclude management transactions as with the independent mining model. However, it shifts the balance of power slightly towards the currency administrator by allowing them to convey a small financial advantage to preferred miners. This risk can be arbitrarily diminished through making certain permanent policy settings.

The theoretical architecture defined in [1] proposed modifying Bitcoin for its implementation. The original Bitcoin whitepaper is available at [2] while detailed explanations can be found in [3], [4], and [5]. Of import to this work is that Bitcoin transfers coins using transactions. The coins are not stored in user accounts but are linked to the transactions themselves. Thus, each transaction has one or more inputs (Vin fields) that bring unspent coins into the transaction and one or more outputs (Vout fields) that declare who can next spend those coin outputs. As shown in Figure 2, a Vin field from some transaction x brings in an unspent Vout field from some transaction y. Figure 3 shows the format of a Bitcoin transaction.

III. Discovered Vulnerabilities and Security Issues

We discovered vulnerabilities and security issues in the theoretical architecture that needed to be mitigated in order to implement the prototype system. The vulnerabilities enabled violations of the balance of power, replay attacks, and attacks against miners. The security issues included improper scoping of manager and law enforcement powers as well as insecure bootstrapping for establishing cryptocurrency policy.

A. Preserving the Balance of Power

The research in [1] contains a 'dependent mining model' where the manager can specify that x number of management transactions must be included within each interval of y blocks. One can set x and y through issuing policy transactions. The idea is that this model forces the miners to periodically include management transactions.

However, we have discovered a vulnerability in which the manager can use this feature to take over the blockchain. The manager can initially set y to be high and wait for the community to fully adopt and use the cryptocurrency. Once a significant amount of value has been invested in the cryptocurrency, the manager can issue a policy transaction changing y to be very low. The manager then could, for example, require management transactions to be issued with every block and only send those management transactions to miners whom they favor or control. The miners receiving those transactions would then not propagate them to other miners, preventing the other miners from mining any blocks (since per policy all blocks would have to contain a management transaction). This way, only miners that the manager favored or controlled could publish blocks and the manager could effectively take over the blockchain with effects similar to that of a 51 % attack [8].

Our mitigation is to simply limit how tightly a manager can set y. If the specification and developed code reject policy transactions that set y values below some threshold, then the manager is prevented from using this method to take control of the blockchain. The manager could also voluntarily set a minimum threshold for these values using permanent policy transactions issued by the root manager node in order to create public confidence in the cryptocurrency. Even with minimums set, it should be noted that the manager can still implement this attack periodically, favoring their own miners every y blocks if they refuse to issue management transactions in the intervening blocks. This would give a periodic financial advantage to manager favored miners but would be highly visible to the community and would not result in the manager controlling the blockchain. To minimize the impact of this residual attack possibility, y should be required to be high enough to make the financial advantage minuscule. An alternative is to use the independent mining model discussed in II, but this opens up the possibility for the miners to revolt against the manager.

B. Preventing Replay Attacks

The research in [1] modifies the Bitcoin transactions to support roles because the architecture requires that all transactions include roles. They are brought into the transaction using a modified Vin field; in Bitcoin Vin fields are only used to bring coin into a transaction. Both uses of the Vin field use the same cryptographic protections and one would assume that they would inherit the same security properties. However, this is not the case and it results in a vulnerability in the architecture.

Since roles are spent like coins but never get used up (since you don't lose a role through using it), they can be spent an infinite number of times. This means that transactions that use a role might be able to be replayed. For the typical transactions also transferring coin (e.g., to pay a transaction fee), this is not a problem as the replayed transaction will be rejected because the coin would already have been spent. However, if the transaction does not involve coin it could be replayed. This might happen if the manager owns miners servers and issues management transactions without transaction fees with the intention that their miners will publish them. In this case, there would be no barriers to performing a replay attack. This might result in a situation where law enforcement unlocks an account but can never securely lock it again because the original unlocking transaction can be replayed by anyone.

There are several possible solutions. One approach is to require that all transactions pay some transaction fee while requiring transaction signatures to sign the entire transaction. In our attempt to modify Bitcoin as little as possible, our approach was to change the theoretical model to truly spend roles as if they were coin; once spent they can't be spent again. However, whenever we spend a role by including it in a Vin field we also re-create the same role in one of the Vout

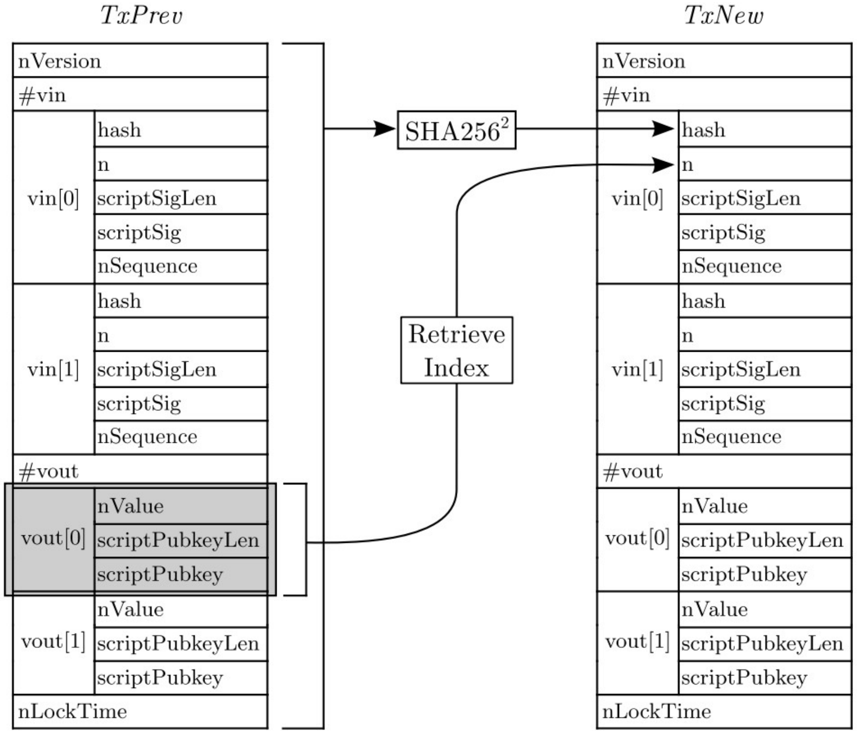
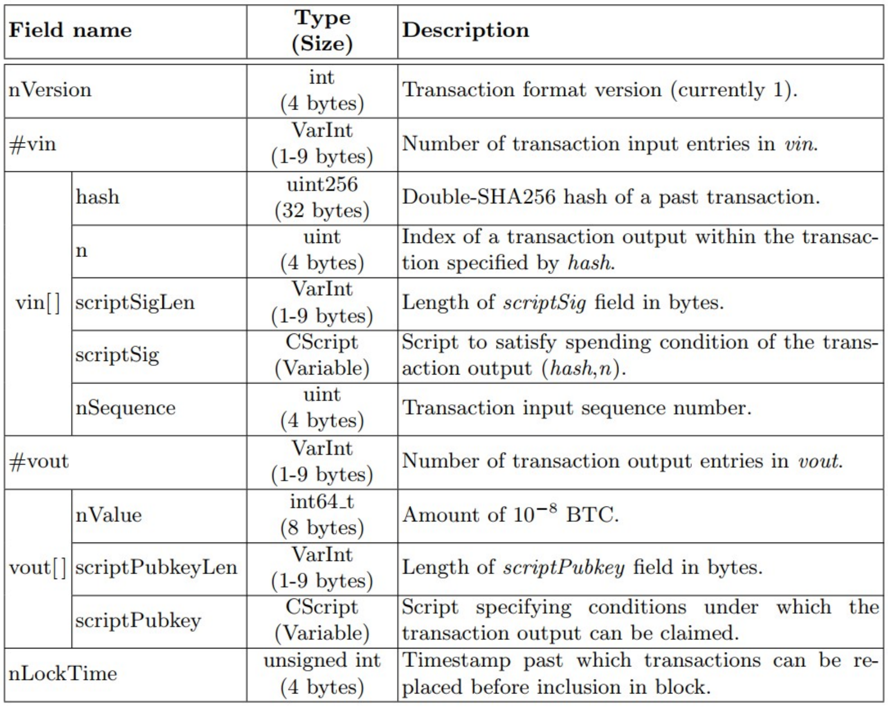


Fig. 2. Bitcoin Vin[] Field Reference to a Previous Transaction (copied from [5]).



Field name		Type (Size)	Description
nVersion		int (4 bytes)	Transaction format version (currently 1).
#vin		VarInt (1-9 bytes)	Number of transaction input entries in *vin*.
vin[]	hash	uint256 (32 bytes)	Double-SHA256 hash of a past transaction.
	n	uint (4 bytes)	Index of a transaction output within the transaction specified by *hash*.
	scriptSigLen	VarInt (1-9 bytes)	Length of *scriptSig* field in bytes.
	scriptSig	CScript (Variable)	Script to satisfy spending condition of the transaction output (*hash*,*n*).
	nSequence	uint (4 bytes)	Transaction input sequence number.
#vout		VarInt (1-9 bytes)	Number of transaction output entries in *vout*.
vout[]	nValue	int64_t (8 bytes)	Amount of 10^{-8} BTC.
	scriptPubkeyLen	VarInt (1-9 bytes)	Length of *scriptPubkey* field in bytes.
	scriptPubkey	CScript (Variable)	Script specifying conditions under which the transaction output can be claimed.
nLockTime		unsigned int (4 bytes)	Timestamp past which transactions can be replaced before inclusion in block.

Fig. 3. Bitcoin Transaction Format for Sending Bitcoin (BTC), copied from [5].

fields. The effect is that an account keeps a role when it is spent but the transaction containing the active version of their role can change. Probably the most elegant approach would be to implement the architecture using a cryptocurrency with an accounts based model so that roles are not stored within transactions, but instead within a record associated with each account (discussed more below).

C. Preventing Managers from Attacking Miners

In [1] all accounts must have the U role for them to receive or spend coin. The purpose is to force all participants in the cryptocurrency to be identity proofed by an account manager in order to receive the U role. This in turn supports 'know your customer' laws, which have been a challenge for most cryptocurrencies [9]. However, this also creates a vulnerability. The manager could keep track of the accounts receiving block rewards and remove the U role from those accounts (thus freezing the funds). The public miners would then have no financial incentive to mine and then the manager's own mining servers could take over the majority of mining. This would give the manager the ability to launch a 51 % attack [8] and to a large degree control the blockchain.

Our solution is to enable miners to deposit block rewards into any account, regardless of whether or not it has been registered in the system or has any roles. Also, we handle the coin from these coinbase transactions (the mining reward transactions) specially such that it can be sent without the owning account needing the U role. This prevents the currency administrator from freezing the mining reward coinbase funds. However, once coinbase coin is sent away from the original account it becomes normal coin subject to the normal requirements (it can't be spent without the associated account having the U role).

D. Scoping Law Enforcement Powers

In [1] law enforcement powers are both too limited and too relaxed. They are too limited in that law enforcement can only lock accounts through removing the U label. Law enforcement nodes can't prevent an account using its other roles (M, C, A, or L). This is a major issue in the event that an account is stolen. On the other hand, law enforcement powers are too relaxed in that law enforcement nodes can effect any node higher in the account hierarchy tree regardless of whether or not it is on the same branch. This effectively gives law enforcement nodes a global reach (which is especially problematic if a law enforcement node is compromised).

Our solution is to reflect account locking not through the removal of the U role but by setting a locked flag. We use one of the unused bits in the nValue field for role change mode to set this flag. If the flag is set it temporarily disables all roles, not just the U role. This stops all activity by the targeted account, giving law enforcement the powers it needs to freeze stolen accounts. At the same time, we put additional restrictions on law enforcement nodes by only giving them authority over nodes farther from the root on the same branch of the node hierarchy tree. More precisely, we define the scope of control of a law enforcement node by traversing backwards until the first node is found with the manager role and then by performing a breadth first search to reveal all nodes within scope. This enables law enforcement nodes to 'hang' off of manager nodes in the tree (they don't have to be inline on each branch).

E. Management Node Powers

In [1] management nodes also had powers that were too relaxed. They were required to have any role that they would want to grant. This resulted in management nodes having powers that they had no intention of using. Also, their scope of control was the same as law enforcement giving each M node low down in the hierarchy tree an almost global reach.

Our solution was to limit their scope to nodes reachable by breadth first search and to limit management nodes to only having the M role. However, in our approach management nodes can add any role to other nodes. This gives more power to a manager node (which might be seen as decreasing security) but it limits that power to a more narrow scope creating what we believe is a rational compromise.

F. Policy Bootstrapping

In [1], it is not stated how the initial policy is defined for an instantiated cryptocurrency. It is implied that some configuration file, apart from the blockchain, must exist that provides the original parameter settings. These settings may or may not then be subsequently overridden through policy transactions on the blockchain. The result may be that some policy is defined on the blockchain and some through an original configuration file. Given that the configuration file wouldn't have the same cryptographic protections as blockchain transactions, the distributor of the node software for maintaining the blockchain could conceivably change policy using software updates through modifying the configuration file.

Our solution is to eliminate the need for the unsecured initial configuration file. We do this by specifying that all policy is initially defined as permissive as possible. We then require that all policy parameters be defined explicitly on the blockchain within the first x blocks (as defined in the full node software distribution). Thus early in the blockchain, ideally prior to it being released publicly, the manager will have to explicitly record all possible policy parameters within cryptographically secured blocks.

We also discovered that the original root management node should not be used to set the initial policy (except for policy settings intended to be permanent). This is because, per [1], management nodes closer to the root are more authoritative; any root manager node policy decisions will prevent any other management node from changing that policy. Also, the root management node account ideally should never be used after the initial few blocks and its keys should be physically stored in a vault to eliminate the possibility of it being compromised. Thus, if the root node is used to set policy it should only be used to set permanent policy that, by design, will never be changed.

IV. Architectural Challenges

Apart from mitigating vulnerabilities in the original design, we encountered several architectural challenges where it was not practical or even possible to directly implement the theoretical architecture. In this section we describe the primary challenges, how we modified the theoretical design to overcome them, and how we implemented those changes.

A. Dual Signature Requirements for Coin Transfer Transactions

In [1] an account must have the U role to both spend and receive coin. It specifies that these roles must be brought into each coin transfer transaction using two separate Vin fields. However, this requires both the sender and receiver to sign the transaction which would require off blockchain coordination and some unspecified infrastructure to support this.

This could be resolved by including the coin transfer recipient only in the Vout field (not the Vin) and requiring full nodes to check the U role on the account listed in the Vout field (without explicitly bringing it into the transaction using a Vin field), at the cost of additional tracking overhead. Our mitigation was to only require the U role for spending coin. Any account then can receive coin, but may not be able to spend it. This results in only a single account needing to sign coin transfer transactions and eliminates additional overhead.

B. Node Movement

In [1] there is no mention of how accounts can change position within the node hierarchy graph once they have been created. This is necessary, for example, for users that want to use different account managers. Besides moving nodes, edges in the graph may need to be moved in order to cut out compromised nodes but leave the rest of the node hierarchy intact.

To implement the needed functionality, we created the idea that if a node adds roles to an account that has no roles, this creates an edge in the node hierarchy graph from the node adding the roles to the node representing the account gaining the roles. If an edge already existed to the node gaining the roles (which would happen if an account received roles and then deleted them), the prior edge will be deleted in order to preserve the required tree structure.

To prepare a node to be moved, the relevant account can unilaterally remove its own roles or else a manager whose scope covers the node can remove the roles. Using this paradigm, nodes can be moved around the node hierarchy tree. It also doesn't require explicitly coding edge creation and deletion within the modified Bitcoin protocol, which would have been unnecessarily complicated. A drawback is that node movement requires a two step process: one transaction to remove roles and another to add them back in (thus removing the old edge and creating the new edge). In our future work we will design a format where a single transaction does this atomically. Complicating this may be the need for dually signed transactions to prevent security violations (which we are trying to avoid, see section IV-A). Our current two step approach ensures that the role removal, node movement, and edge addition only happens through transactions issued by nodes authorized to perform those activities.

C. Determining Transaction Types

The theoretical architecture in [1] uses the most significant bits of an nValue field to determine the type of transaction being processed: role change, policy change, or coin transfer. The nValue fields, in the original Bitcoin, specify the amount of coin to be spent. Using the leftmost bits as control bits is conceivably risky because a bug in the code might interpret the leftmost control bits as value bits for moving or create large amounts of coin. More problematic though is that the Bitcoin implementation uses the leftmost bit of the nValue field as a signed bit.

For these reasons, we chose to deviate from [1] and not use the leftmost bits of the nValue field to determine the type of transaction. Instead, we determined the type of transaction using the transaction version number; this then determines how the nValue fields within a transaction are handled. We created three transaction version numbers, each of which correspond to the three different modes for evaluating nValue fields (role change, policy change, and coin transfer). Lastly, we also changed to using the nValue low order bits for specifying roles and policy change types in case those nValue fields ever got interpreted as coin transfer fields through some bug or attack. This would then limit the damage done by having fewer coin inadvertently transferred or created.

D. Transaction Fees

Since we determine transaction type (role change, policy change, or coin transfer) through the transaction version number, it means that the mode of all the nValue fields in the Vout fields are determined by that number. However, it is usually necessary to pay a transaction fee for most transactions and there is usually change that must be sent back to the sender. This is not possible then for the role and policy change transactions because the nValue fields of the Vout fields change roles/policies; they don't send coin as in the original Bitcoin specification. We solved this simply by specifying which Vout field is always the change sent back to the originator of the transaction (which may be 0 coin on occasion).

V. Developed Prototype

Our prototype was developed publicly through Github and is available within the project 'usnistgov/managed-cryptocurrencies-bitcoin'. We built our prototype through forking and modifying the C++ Bitcoin codebase available on Github at 'bitcoin/bitcoin'.

For flexibility, efficiency, and portability we ran our modified bitcoin peer-to-peer network for development and testing on a local virtualized environment. For our testing, we thus had a single virtual machine (VM) executing the entire distributed Bitcoin network. We used the Vagrant virtual machine manager with Virtualbox as the VM provider. Within the VM, we used the Docker Engine to run a set of containers to represent the nodes on the modified Bitcoin network. This enabled us to simultaneously run five Bitcoin miners within a single VM to maintain our test blockchain. Note that we artificially reduced the mining difficulty to enable quick block production for testing and demonstration purposes. Lastly, we used the GraphViz library to enable us to visualize the node hierarchy tree. To make access control decisions for role and policy change transactions, it was inefficient to look up individual node roles using the tree. Thus, we separately maintained an associative array mapping node names to a list of their roles. The tree was only necessary for determining the scope of control of one node over others (e.g., for the law enforcement and manager nodes).

An example output tree is shown in Figure 4. Within each node in parenthesis is listed the roles activated for that node and its state (locked or unlocked). The labels are deciphered as follows: M-manager, C-central banker, L-law enforcement, U-registered user, A-account manager, D-disabled account) Node 0 is the root node created in the genesis block. It should normally never be used directly for security reasons and so Node 1 was created as the 'active' manager. Node 3 is the central banker; it could have hung off of Node 1 but it was useful for our example to have it as a child under Node 0. Node 2 is law enforcement with the scope of all that is reachable from Node 1 (all nodes except 0, 3, and 11). Nodes 4 and 5 are account managers. Node 6 is a user account that has been disabled by law enforcement. Nodes 7, 8, and 10 are ordinary users. Node 9 is a node who has had all its roles removed (either done by Node 9 itself, its account manager Node 5, or one of the manager nodes 0 or 1). This might have been done because Node 9 was compromised or because it is being prepared to move to another part of the tree under a different account manager. Node 11 is a node that has been active in the cryptocurrency but has no roles and has never had any roles (due to their being no edge to it). It represents an account created by a miner to store coinbase coins, that can be spent without needing any roles.

VI. Related Work

To our knowledge, [1] is the only work proposing a managed cryptocurrency that has a balance of power where the public can hold the manager accountable. There have been many government cryptocurrencies proposed but these differ in that they are often not managed, don't use roles, or don't have a balance of power.

Multichain [10] is a system that might appear to be similar in that it contains management features. However, Multichain enables a permissioned chain where what is managed is which entities have the privilege of mining. This is opposite of our prototype that enables open mining. That said, we may explore modifying Multichain to implement [1] while leveraging a permissioned chain whose membership is defined by the current members (not the manager).

There are many government cryptocurrencies proposed and in development (for example [11], more citations are in [1]). However, none of these have yet come to fruition except the Venezuelan Petro [12], which to our knowledge is the only existing government issued cryptocurrency.

There is research proposing a Fedcoin [13], a cryptocurrency that would support central banks with a permissioned blockchain that complies with 'know your customer' laws [9]. It is based on RS—Coin [14], one of many cryptocurrencies advertised to support central and commercial banks with international transaction handling. Others argue that central banks don't need a cryptocurrency, but instead a new form of electronic money [15]. There are also concerns with the amount of power a government could leverage through creating a Fedcoin [16].

VII. Future Work

There are two major changes to be made in future iterations of the implementation: using an account model and better handling of compromised nodes.

A. *Using an Account Model*

Bitcoin uses an unspent transaction output (UTXO) model. Coin is not stored within user accounts but within the transactions themselves. All transactions have outputs (representing coin) and any unspent output may be spent by another transaction. Who may spend a given output is determined by a script that usually specifies the public key of a particular account. There is no single data structure on the blockchain that shows the coin associated with a particular account.

This works well for Bitcoin, but immediately became awkward for the implementation of our managed cryptocurrency prototype. In the theoretical architecture, accounts have roles that specify their privileges in the system and these roles are specified in nValue fields. Without a central data structure for each account, the roles had to be treated like coin and be spent repeatedly as an account used those roles. In our system, an account's roles are transaction outputs and the active copy (the one that hasn't yet been spent) is temporarily in one particular transaction. We simplified this, compared to the theoretical architecture, by requiring that any role additions and removals repeat the remaining roles. Thus, all of an account's roles are always designated within a single transaction, not spread out among many transactions as would have occurred through a direct implementation of the theoretical architecture.

Our future approach will be to implement the system through forking cryptocurrency code that uses an account model instead of an UTXO model. This is possible because the theoretical architecture is not tied to any particular cryptocurrency. A likely candidate replacement cryptocurrency would be Ethereum due to its maturity, but this choice would bring in the added complexity of a codebase that supports smart contracts. A mature Bitcoin-like cryptocurrency without smart contract capabilities that uses an account model might be better suited.

B. *Handling Compromised Nodes*

In section III-D we expand the law enforcement powers to disable all the roles of an account to handle the case where

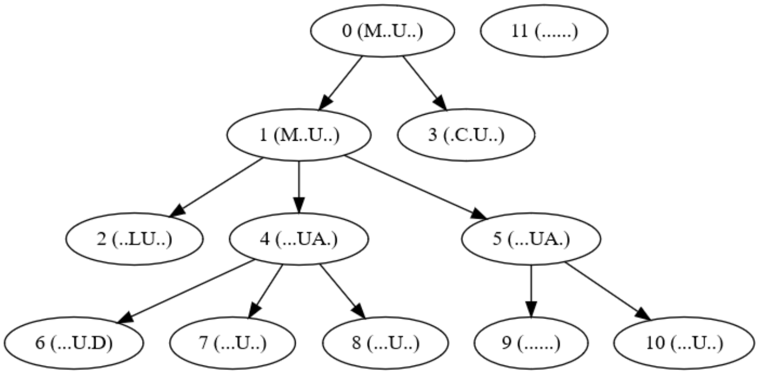


Fig. 4. Example Output Showing a Node Hierarchy.

a node is compromised (in [1] only the ability to send and receive coin was disabled). However, this does not allow the compromised node to be recovered. To do this, we propose that all nodes should have two sets of cryptographic key pairs. The first set is used for the daily signing of transactions for the associated account. The second set is stored offline and is used only to replace the first set. This enables account owners to unilaterally re-establish control over their accounts without having to involve a manager node (one with the M or A role). However, it will require the development and implementation of a new transaction type to enable this resetting of the first key pair.

VIII. Conclusion

The theoretical managed cryptocurrency architecture proposed in [1] can be efficiently developed from an existing cryptocurrency codebase and deployed (despite the many implementation issues that had to be overcome). An important result of this is that we have shown that the novel balance of power concept, whereby a manager and public miners jointly control a cryptocurrency, is a feasible mechanism to be explored for future cryptocurrencies. Another result of our work is to show the practicability of adding roles to cryptocurrency accounts and the capabilities that can be achieved through these roles (in particular for mimicking fiat currency mechanisms). Lastly, we note that building such a protocol native managed cryptocurrency within a blockchain platform itself was non-trivial but we showed that it could be accomplished with only a modest cost in programming effort.

In summary, we have shown that the theoretical system in [1] can be implemented in such a way as to not just leverage many of the strengths of modern cryptocurrencies, but also leverage the capabilities of traditional fiat currencies. While this goes against the goals and directions of most cryptocurrency efforts which are promoting greater privacy and autonomy from managing institutions, this result may be useful for large institutions (e.g., governments) investigating future electronic currency approaches. We do not necessarily believe that the architecture in [1] provides the answer for such a use case, but it and our applied research in this work may open up new research directions to better support large institutions issuing their own managed cryptocurrencies.

References

[1] P. Mell, “Managed blockchain based cryptocurrencies with consensus enforced rules and transparency,” in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2018, pp. 1287–1296.

[2] S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[3] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten, “Sok: Research perspectives and challenges for bitcoin and cryptocurrencies,” in *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, 2015, pp. 104–121.

[4] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press, 2016.

[5] K. Okupski, “Bitcoin developer reference,” 2016. [Online]. Available: https://lopp.net/pdf/Bitcoin_Developer_Reference.pdf

[6] “bitcoinwiki protocol documentation,” accessed: 2017-12-29. [Online]. Available: https://en.bitcoin.it/wiki/Protocol_documentation

[7] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, “Where is current research on blockchain technology? a systematic review,” *PloS one*, vol. 11, no. 10, p. e0163477, 2016.

[8] S. Barber, X. Boyen, E. Shi, and E. Uzun, “Bitter to betterhow to make bitcoin a better currency,” in *International Conference on Financial Cryptography and Data Security*. Springer, 2012, pp. 399–414.

[9] M. Staples, S. Chen, S. Falamaki, A. Ponomarev, P. Rimba, A. Tran, I. Weber, X. Xu, and J. Zhu, “Risks and opportunities for systems using blockchain and smart contracts,” 2017. [Online]. Available: https://publications.csiro.au/rpr/download?pid=csiro:EP175103dsid=DS2

[10] G. Greenspan, “Multichain private blockchainwhite paper,” 2015. [Online]. Available: https://www.multichain.com/download/MultiChain-White-Paper.pdf

[11] L. Coleman. An inside look at chinas government controlled cryptocurrency project. [Online]. Available: https://www.ccn.com/an-inside-look-at-chinas-government-controlled-cryptocurrency-project

[12] D. B. Alexandra Ulmer, “Enter the ’petro’: Venezuela to launch oil-backed cryptocurrency,” Reuters, Dec. 2017.

[13] S. Gupta, P. Lauppe, and S. Ravishankar, “A blockchain-backed central bank cryptocurrency,” 2017. [Online]. Available: https://zoo.cs.yale.edu/classes/cs490/16-17b/gupta.sahil.sg687

[14] G. Danezis and S. Meiklejohn, “Centrally banked cryptocurrencies,” *arXiv preprint arXiv:1505.06895*, 2015.

[15] A. Berentsen and F. Schar, “The case for central bank electronic money and the non-case for central bank cryptocurrencies,” 2018. [Online]. Available: https://doi.org/10.20955/r.2018.97-106

[16] T. Aube. The terrifying future of fedcoin. [Online]. Available: https://hackernoon.com/the-terrifying-future-of-fedcoin-ddcbef2b9592

TMPS: Ticket-Mediated Password Strengthening

Abstract

We introduce the notion of Ticket-Mediated Password Strengthening (TMPS), a technique for allowing users to derive keys from passwords while imposing a strict limit on the number of guesses of their password any attacker can make, and strongly protecting the users' privacy. We describe the security requirements of TMPS, and then a set of efficient and practical protocols to implement a TMPS scheme, requiring only hash functions, CCA2-secure encryption, and blind signatures. We provide several variant protocols, including an offline symmetric-only protocol that uses a local trusted computing environment, and online variants that avoid the need for blind signatures in favor of group signatures or stronger trust assumptions. We formalize the security of our scheme by defining an ideal functionality in the Universal Composability (UC) framework, and by providing game-based definitions of security. We prove that our protocol realizes the ideal functionality in the random oracle model (ROM) under adaptive corruptions with erasures, and prove that security w.r.t. the ideal/real definition implies security w.r.t. the game-based definitions.

1 Introduction

In many real-world cryptosystems, the user's password is the greatest practical weakness. Commonly, the user enters a memorized password or passphrase, which is processed using a password-based key derivation function, to get a symmetric key. Knowledge of this key allows the attacker to completely violate the security of a system he has physically compromised–to read the user's files, sign arbitrary things with her private key, spend her bitcoins, etc.

Deriving a key from a password is an old, well-studied problem [18, 25, 30]. The practical difficulty comes from the ability of an attacker to run a parallel password search. After stealing the user's device, the attacker is able to run an offline attack, perhaps trying billions of different password guesses per second until the user's password is found. Few users can memorize a password capable of withstanding such an attack for long.

In this paper, we propose a novel method to approach this problem, called Ticket-Mediated Password Strengthening (TMPS). Informally, this works as follows: When the user wants to produce a new password-derived key, she creates and stores some local data, and runs a protocol with a server to produce a set of t *tickets*. Later, when she wants to unlock this key using her password, she runs another protocol with the server, providing (and discarding) one of these tickets. The password can only be used to unlock the user's key with the server's help, and the server will not provide this help without a ticket that has never before been used.

The result is that the user can establish a hard limit on the number of possible guesses of the password any attacker can make–if she has only 100 tickets on her device, then an attacker who compromises the device can never try more than 100 guesses of her password.

Our approach provides strong privacy guarantees for the user w.r.t. the server, and unlike many proposed password-hardening schemes in the literature, is focused on user-level key derivation problems, rather than on online authentication to some web service. We also provide a mechanism to allow the server to control which users it provides services to, without violating the users' privacy.

Among other advantages, our scheme gives users a security metric that is human-meaningful–the maximum number of guesses the attacker can ever have against their password. Hardness parameters of password hashes, or entropy estimates of a password, are meaningful only to security experts; the maximum number of attacker

Kelsey, John M.; Dachman-Soled, Dana; Sonmez Turan, Meltem; Mishra, Sweta. "TMPS: Ticket-Mediated Password Strengthening." Paper presented at The Cryptographer's Track of the RSA Conference (CT-RSA 2020), San Francisco, CA, US. February 24, 2020 - February 28, 2020.

guesses that will be allowed is much easier to understand. On the other hand, our scheme imposes the need to be online in order to unlock a key secured by a password. (Though we provide variant schemes which can be used offline.)

1.1 Our Results

- We introduce the notion of Ticket-Mediated Password Strengthening (TMPS), a mechanism for allowing users to derive keys from passwords while imposing a strict limit on the number of guesses of their password any attacker can make, and strongly protecting the users' privacy.
- We formalize the security requirements of our new notion of TMPS, both by introducing game-based definitions (See Appendix B) as well as defining a corresponding ideal functionality in the Universal Composability (UC) framework (See Section 5).
- We present efficient protocols realizing our new notion. Our basic protocol requires only hash functions, CCA2-secure encryption, and blind signatures (See Section 4).
- We prove that our protocol UC-realizes the aforementioned ideal functionality in the random oracle model (ROM) under adaptive corruptions with erasures (See Appendix C) and prove that security w.r.t. the Ideal/Real definition implies security w.r.t. the game-based definitions (See Appendix B).
- We present several variants of our protocol, including an offline version of our protocol using a local hardware security module (HSM) or trusted execution environment, and variants that make use of group signatures, proofs of work, or weaker security assumptions to ensure user privacy while still preventing overuse of server resources (See Section 6).
- Finally, we discuss efficient implementations and performance, and consider some questions left open by this research, in Section 7.

1.2 Related Work

A long list of work (e.g., [5,6,31]) focuses on Password Authenticated Key Exchange (PAKE) protocols where the user and server share a password to securely establish a session key. These protocols have some similarities to our scheme (in particular the use of a trusted server), but their details (key generation) and usage are quite different. Typically, the PAKE protocols are vulnerable to offline dictionary attacks after server compromise which has been tackled in recent proposals such as Qpaque protocol [16]. Another related concept known as Password Protected Secret Sharing (PPSS) as described in [15], requires many servers to hold some share of a key; a secret sharing scheme is used to reconstruct them, secured by a password.

Many significant works on password security focuses on password-based authentication systems. In [24], Mani describes a scheme that uses a server to assist in password hashing, but without any concern for user privacy–the goal in that scheme was to harden the password file by incorporating a PRF computed on a single-purpose machine. Similarly, [2,26,27] describe a scheme with a separately-stored secret key in a crypto-server to strengthen password hashing, an informal description of the concept of password hardening, later formally defined in [12,22,28]. Our scheme differs in goal; however, is related to *password hardening* in using a separate server or device, in order to limit (online) brute-force attacks.

Another significant proposal by Lai et al. [21] defines a *password hardening encryption* scheme which provides substantial protection from password-cracking attacks by rate-limiting password cracking attempts assuming the crypto server is not compromised. Our scheme differs in many ways, most notably in the use of tickets for decryption, and in the assurance of user privacy. Also, our scheme focuses on the setting of password-based key derivation on the user's device, rather than on a server the user trusts with her data.

Still another line of work introduces the notion of password-based threshold authentication [1] for token-based authentication in single sign-on setting–in their scheme, any subset of ℓ of n servers participate in verifying the user's password and generating a token.

There is also a rich literature in server-assisted computation [3,13,19,20] which preserves the user's privacy. Our scheme differs from most of this work in that we are not trying to offload computational work to the server, we are just using a server to limit the number of times some computation may be done. In our proposal, we use standard algorithms, and provide a great deal more flexibility to choose the underlying cryptographic functions than is available in these schemes.

2 Preliminaries

2.1 Notation

Let $k \in \mathbb{N}$. The set of bitstrings of length k is denoted as $\{0,1\}^k$. The concatenation of two bitstrings x and y is denoted by $x \parallel y$. The exclusive-OR of two bitstrings x

Kelsey, John M.; Dachman-Soled, Dana; Sonmez Turan, Meltem; Mishra, Sweta. "TMPS: Ticket-Mediated Password Strengthening." Paper presented at The Cryptographer's Track of the RSA Conference (CT-RSA 2020), San Francisco, CA, US. February 24, 2020 - February 28, 2020.

and y of same length is denoted as $x \oplus y$. We let b^k denote the string with k successive repetitions of bit b. If $\mathcal{X}$ is a set, we let $x \leftarrow_{\$} \mathcal{X}$ denote sampling a uniformly random element x from $\mathcal{X}$. The security parameter is denoted by $n \in \mathbb{N}$. Unless otherwise specified, we assume all symmetric keys and hash outputs to be n bits in length.

2.2 Underlying Primitives and Functions

We use the following primitives in our protocols:

- $\texttt{HASH}(X)$: The cryptographic hash of input X.
- $\texttt{HMAC}(K,X)$: The HMAC of X under key K.
- $\texttt{PH}(S,P)$: Hash of the password P using salt S.
- $\texttt{KDF}(K,D,\ell)$: ℓ-bit key derived from the secret value K and public value D.
- Π_{ENC} := (GEN, ENC, DEC): An encryption system where $\texttt{ENC}(K,X)$ is encryption of plaintext X under the key K, and $\texttt{DEC}(K,Y)$ is decryption of ciphertext Y under the key K.
- Π_{BSIG} := (GEN, BLIND, UBLIND, SIGN, BVERIFY): A 2-move blind signature scheme where
  - $M^* \leftarrow \texttt{BLIND}(M)$: The user blinds the message M to obtain M^* and sends to the signer.
  - $\sigma^* \leftarrow \texttt{SIGN}_{SK}(M^*)$: The signer outputs a signature σ^* on input of message M^* and private key SK and sends to the user.
  - $F \leftarrow \texttt{UBLIND}(\sigma^*)$: The user unblinds the signature σ^* to obtain F. Note that the user inputs additional private state to the UBLIND algorithm, which we leave implicit.
  - $\texttt{BVERIFY}_{PK}(M,F)$: Verification of signature F on message M under public key PK as valid/invalid.

Next, we define two internal functions: $\texttt{VE}(D,K_P)$ provides verifiable encryption of K_P with D and $\texttt{DV}(D,Z)$ decrypts K_P after checking the correctness of D. Both functions assume that D, K_P and hash outputs are n bits long.

Algorithm 1 Verifiably encrypt K_P with D.

```
function VE($D, K_P$)
    $Z \leftarrow \texttt{HASH}(0 \parallel D) \parallel (\texttt{HASH}(1 \parallel D) \oplus K_P)$
    return($Z$)
```

Algorithm 2 Verifiably decrypt Z with D.

```
function DV($D, Z$)
    $X \leftarrow Z_{0\ldots n-1}$
    $Y \leftarrow Z_{n\ldots 2n-1}$
    $X^* = \texttt{HASH}(0 \parallel D)$
    if $X == X^*$ then
        return($\texttt{HASH}(1 \parallel D) \oplus Y$)
    else
        return($\perp$)
```

3 Ticket-Mediated Password Strengthening

In *Ticket-Mediated Password Strengthening*, or TMPS, users first interact with a server to get a set of *tickets*. Each ticket entitles the user to assistance from the server with one attempt to unlock a master secret (called the *payload key*) using a password. Later, users (or anyone else with access to the tickets) may use the tickets to attempt to unlock the payload key using the password.

TMPS requires a setup phase, and two protocols: REQUEST and UNLOCK. During setup, the server establishes public encryption and signing keys and makes them available to its users.

In order to get tickets, the user chooses a payload key and a password, and runs the REQUEST protocol with the server, requesting t tickets. If the protocol terminates successfully, the user ends up with t tickets, each of which entitles her to one run of the UNLOCK protocol.

In order to use a password to unlock the payload key, the user must consume a ticket–she runs the UNLOCK protocol with the server, passing the server some information from the ticket and some information derived from her ticket and her password. When the protocol runs successfully, the user recovers the payload key.

The security requirements of a TMPS scheme are:

1. REQUEST
   (a) The server learns nothing about the password or payload key from the REQUEST protocol.
   (b) There is no way to get a ticket the server will accept, except by running the REQUEST protocol.
   (c) Each ticket is generated for a specific password and payload key; tickets generated for one password and payload key give no help in unlocking or learning any other password or payload key.
2. UNLOCK

(a) An UNLOCK run will be successful (it will return the correct K_P) if and only if:

i. This ticket came from a successful run of the REQUEST protocol.
ii. This ticket has never been used in another UNLOCK call.
iii. The same password used to request the ticket is used to UNLOCK it.

(b) From an unsuccessful run of the UNLOCK protocol, the user gains no information about the payload key.

(c) From an unsuccessful run of the UNLOCK protocol, the user learns (at most) that the password used to run the protocol was incorrect.

(d) The server learns nothing about the payload key or password from the UNLOCK protocol.

(e) The server learns nothing about which user ran the UNLOCK protocol with it at any given time.

Note that these requirements don't describe the generation of the payload key or the selection of the password. If the payload key is known or easily guessed, then TMPS can do nothing to improve the situation. In any real-world use, the payload key should be generated using a high-quality cryptographic random number generator.

The strength of the password matters for the security of ticket-mediated password strengthening, but in a very limited way–each run of UNLOCK consumes one ticket and allows the user to check one guess of the password. An attacker given N equally-likely passwords and t tickets thus has at most a t/N probability of successfully learning the password.

3.1 Discussion

The usual way password-based key derivation fails is that an offline attacker tries a huge number of candidate passwords, until he finally happens upon the user's password. He then derives the same key as the user derived, and may decrypt her files. A TMPS scheme avoids this attack by requiring the involvement of the server in each password guess, and (more importantly) by limiting the number of guesses that will ever be allowed. If the user of a TMPS scheme requests only 100 tickets from the server, then an attacker who compromises her machine and learns the tickets will never get more than 100 guesses of her password. If he cannot guess the password in his first 100 guesses, then he will never learn either the password or the payload key. Even if he is given the correct password after he has used up all the tickets, he cannot use that password to learn anything about the payload key.

The security of a TMPS scheme relies on the server being unwilling to allow anyone to reuse a ticket, and the inability of anyone to unlock a payload key with a password *without* running the UNLOCK protocol with a server, and consuming a fresh ticket in the process.

A corrupt server can weaken the security of TMPS, but only in limited ways. It cannot learn anything about the password or payload key. It cannot determine which user is associated with which ticket, or link REQUEST and UNLOCK runs. But it *can* enable an attacker who has already compromised a user's tickets to reuse those tickets as many times as he likes.

4 The Basic Protocol

In this section, we describe a set of protocols that implement Ticket-Mediated Password Strengthening in a concrete way. Our protocols require a secure cryptographic hash function, a public key encryption scheme providing CCA2 security[1], and a blind signature scheme. (Note that there are variants that do not need a blind signature scheme described in Section 6.)

A *ticket* gives a user enough information to enlist the server in helping carry out one password-based key derivation. Each ticket contains an *inside* part (which the user retains and does not share with the server) and an *outside* part (which the user sends to the server). The different parts of a ticket are bound together with each other and with the specific password and key derivation being carried out, and can't be used for a different key derivation.

We make two assumptions about this protocol: First, all messages in this protocol take place over an encrypted and authenticated channel. Second, the user somehow demonstrates that he is entitled to be given tickets by the server; we assume the user has already done this before the REQUEST protocol is run. There are many plausible ways this might be done, such as:

1. The user may pay per ticket.
2. The user may demonstrate his membership in some group to whom the server provides this service.
3. The server may simply provide this service for all comers.

The specific method used is outside our scope. However the user demonstrates his authorization to receive tickets, it is very likely to involve revealing her identity. In

[1] An attacker who can alter a ciphertext to get a new valid ciphertext for the same plaintext can attack our scheme.

order to protect the user's privacy from the server, the REQUEST protocol must thus prevent the server linking tickets with this identifying information, or linking tickets issued together.

4.1 Server Setup

The steps given below are done once by the server (though presumably the server rolls over keys from time to time).

- The server establishes an encryption keypair PK_S, SK_S for some algorithm that supports CCA2 security. Server distributes its public key to all users.
- The server establishes a signature keypair PK'_S, SK'_S for some algorithm that supports blind signatures.
- The server establishes a list to store previously-seen tickets.

4.2 REQUEST: Protocol for Requesting Tickets

The basic ticket requesting protocol is illustrated below. The user starts out with a password P and a payload key K_P, and generates t tickets with the assistance of the server.

In order to create a ticket without revealing any identifying information to the server, the user will do the following steps:

1. Randomly generate an n-bit salt S and an n-bit secret value B.
2. Encrypt B using the public encryption key PK_S of the server, producing E.
3. Run a protocol to get a blind signature on E from the server–this is F.
4. Derive a one-time key from the password and the secret B:
$$D \leftarrow \mathtt{HMAC}(B, \mathtt{PH}(S, P))$$
5. Encrypt the payload key under the one-time key: $X \leftarrow \mathtt{VE}(D, K_P)$

The ticket will consist of (S, E, F, Z); the user must irretrievably delete all the intermediate values above. The user repeats the steps t times to get t tickets. Below, we show the REQUEST protocol:

```
Protocol: REQUEST$(P, K_P, t)$:

User                                   Server
for $i = 1 \ldots t$
    $S \leftarrow_{\$} \{0,1\}^n$
    $B \leftarrow_{\$} \{0,1\}^n$
    $E \leftarrow \mathtt{ENC}(PK_S, B)$
    $E^* \leftarrow \mathtt{BLIND}(E)$
                        $\xrightarrow{E^*}$
                                       $\sigma^* \leftarrow \mathtt{SIGN}_{SK'_S}(E^*)$
                        $\xleftarrow{\sigma^*}$
    $F \leftarrow \mathtt{UBLIND}(\sigma^*)$
    $C \leftarrow \mathtt{PH}(S, P)$
    $D \leftarrow \mathtt{HMAC}(B, C)$
    $Z \leftarrow \mathtt{VE}(D, K_P)$
    Forget $B, C, D, E^*, \sigma^*$
    $T_i \leftarrow (S, E, F, Z)$
endfor
return$(T_{1,2,\ldots,t})$
```

At the end of this protocol, the user constructs t tickets she can use to run the UNLOCK protocol. The server, on the other hand, knows only that it has issued t tickets–it knows nothing else about them!

4.3 UNLOCK: Protocol for Unlocking a Ticket

In order to use a ticket along with a password $\hat{P}$ to unlock K_P, the user does the following steps:

1. Hash the password: $\hat{C} \leftarrow \mathtt{PH}(S, \hat{P})$.
2. Send $(E, F, \hat{C})$ to the server.
3. If the signature is invalid or E is being reused, then the server returns $\perp$.
4. Otherwise:
   (a) The server stores E, F as a used ticket.
   (b) $B \leftarrow \mathtt{DEC}(SK_S, E)$
   (c) $D \leftarrow \mathtt{HMAC}(B, \hat{C})$
   (d) The server sends back D.
5. The user tries to decrypt Z with D. If this succeeds, she learns K_P. Otherwise, she learns that $\hat{P}$ was not the right password.

Protocol: $\textsc{Unlock}(S, E, F, Z, \hat{P})$:

```
User                                    Server
$\hat{C} \leftarrow \texttt{PH}(S, \hat{P})$
                  $\xrightarrow{E, F, \hat{C}}$
                                        IF
                                          $E$ fresh AND
                                          $\texttt{VERIFY}_{SK'_S}(E, F)$
                                        THEN
                                          $B \leftarrow \texttt{DEC}(SK_S, E)$
                                          $D \leftarrow \texttt{HMAC}(B, \hat{C})$
                                        ELSE
                                          $D \leftarrow \perp$
                  $\xleftarrow{D}$
$K_P \leftarrow \texttt{DV}(D, Z)$
return$(K_P)$
```

Note that in these two protocols, the server never learns anything about K_P, P, or $\hat{P}$, and has no way of linking a ticket between REQUEST and UNLOCK calls.

5 Security Analysis

In this section, we provide a security analysis and some security proofs for our basic protocol. Our approach comes in three separate parts: First, we define an *ideal functionality* for the system. Second, we prove that our basic protocol is indistinguishable from the ideal functionality in the UC framework. Finally, we provide several game-based security definitions, and prove bounds on an attacker's probability of winning the games when they are interacting with the ideal functionality. These game-based definitions show that the ideal functionality we've defined actually provides the practical security we need from this scheme.

The ideal functionality makes use of a table τ–a key-value database indexed by a ticket T. T can be any n-bit string, or the special values $\perp$ and $*$.

A user calls REQUEST to get a new ticket[2]. We assume a two-sided authenticated and secure channel for REQUEST–the ideal functionality knows the user's identity, and the user knows she is talking with the ideal functionality. Also, REQUEST requires an interaction with the server, in which the server also learns the user's identity. At the end of the REQUEST call, the user either has a valid ticket, or knows she did not get a valid ticket.

Algorithm 3 Ideal Functionality: Initialize and REQUEST

```
function INITIALIZE(SID)
    $\textsc{sid}.\tau \leftarrow \{\}$
function REQUEST$(U, \textsc{sid}, P, K_P)$
    $T \leftarrow_\$ \{0,1\}^n$
    $\textsc{sid}.\tau[T] \leftarrow (P, K_P, \perp)$
    Send to server SID: $(\textsc{sid}, \textsc{Request}, U)$
    if server SID compromised then
        Wait for response $(\textsc{sid}, \textsc{Request}, U, R)$.
    else
        $R \leftarrow 1$
    if $R = 1$ then
        Send to source $U$: $(\textsc{sid}, \textsc{Request}, T)$
    else
        Send to source $U$: $(\textsc{sid}, \textsc{Request}, \perp)$
```

The user makes use of a ticket and a password to recover her payload key with an UNLOCK call. We assume the UNLOCK call is made over a secure channel which is authenticated on one side–the user knows she is talking with the ideal functionality, but the ideal functionality doesn't know who is talking to it. UNLOCK also requires an interaction with the server, in which the server is not told the identity of the user. At the end of the UNLOCK call, the user either learns the payload key associated with the ticket she has used, or receives an error message ($\perp$) and knows the UNLOCK call has failed.

Before stating our theorem, we note that we assume that the protocols for REQUEST and UNLOCK given in Sections 4.2 and 4.3 are executed in a hybrid model, where an ideal functionality for secure, two (resp. one)-sided authenticated channels, $\mathcal{F}_{\mathsf{ac}}$ (resp. $\mathcal{F}_{\mathsf{osac}}$), (see e.g. [8]) is invoked each time a message is sent. We require that the VE scheme used is the one given in Algorithms 1 and 2. We assume three independent random oracles: $H_{\mathsf{pw}}, H_{\mathsf{KD}}, H_{\mathsf{VE}}$. H_{pw} is the password hash. H_{KD} is used to model the HMAC key derivation as a random oracle and H_{VE} is the random oracle for the verifiable encryption scheme given in Algorithms 1, 2.

Theorem 5.1. Under the assumption that Π_{ENC} is a CCA2-secure encryption scheme (see Definition A.5), Π_{BSIG} is a 2-move blind signature scheme (see Definition A.7) and the assumptions listed above, the protocols for SETUP, REQUEST and UNLOCK given in Sections 4.1, 4.2 and 4.3, UC-realize the ideal functionality

[2] The ideal functionality is defined for one ticket, but in our protocol, we define REQUEST to return t tickets at a time. This is equivalent to just rerunning the REQUEST ideal functionality t times.

Algorithm 4 Ideal Functionality: UNLOCK

```
# If ticket and password good, return $K_P$.
# Otherwise, return $\perp$.
function UNLOCK($\text{SID}, T, \hat{P}$)
    if $T \in \text{SID}.\tau$ then
        $(P, K_P, \alpha) \leftarrow \text{SID}.\tau[T]$
    else
    # $\alpha = *$ signals invalid ticket.
        $(P, K_P, \alpha) \leftarrow (\perp, \perp, *)$
        $R \leftarrow 0$
    if $\alpha = \perp$ then
    # Fresh ticket
        $\alpha \leftarrow_{\$} \{0,1\}^n$
        $R \leftarrow 1$
    else
    # Reused or invalid ticket
        $R \leftarrow 0$
    # Server can see whether it's getting invalid,
    # repeated, or fresh ticket.
    Send to server SID: $(\text{SID}, \text{UNLOCK}, \alpha)$
    # If server is NOT compromised, we know R.
    # If server IS compromised, we must ask it
    # how to respond.
    if Server SID compromised then
        Wait for $(\text{SID}, \text{UNLOCK}, R)$  # $R \in \{0,1\}$
    # Send back the right response to the user.
    if $R = 0$ then
    # Server returns $\perp$, no decryption possible.
        Respond to caller: $(\text{SID}, \text{UNLOCK}, \perp)$
    else if $R = 1$ then
    # Server plays straight.
        if $\hat{P} = P$ then
            Respond to caller: $(\text{SID}, \text{UNLOCK}, K_P)$
        else
    # Server returns value, decryption fails.
            Respond to caller: $(\text{SID}, \text{UNLOCK}, \perp)$
```

provided in Algorithms 3 and 4 under adaptive corruptions, with erasures.

We note that our protocols can be generalized to work with multi-round blind signature schemes, and the same security proof goes through.

The proof of this theorem appears in Appendix C.

6 Variants of the Basic Protocol

In this section we discuss some variants and modifications of the basic protocol which may be useful in specific situations.

6.1 Limiting Password Attempts

Ticket-mediated password strengthening permits a user to request a large number of tickets at once, and this may make sense for reasons of efficiency or convenience. However, if the user has chosen a very weak password, it would be helpful to limit any attacker who compromises the user's machine to a very small number of password guesses. For example, many systems have a limit of ten password attempts before locking an account. There is a straightforward way to get this same limit with ticket-mediated password strengthening, even when requesting hundreds or even thousands of tickets at a time.

Suppose the user has successfully created 1000 password-hashing tickets, $T_{0,1,2,\ldots,999}$. Each successful use of a password ticket derives the payload key, K_P. In order to implement a limit of at most ten password guesses, we do the following steps:

1. Setup:
   (a) $K_T \leftarrow \texttt{KDF}(K_P, \text{“ticket encryption”}, n)$
   (b) Using any AEAD scheme, individually encrypt all but ten tickets under the key K_T
2. Each time a ticket is successfully used to unlock K_P
   (a) $K_T \leftarrow \texttt{KDF}(K_P, \text{“ticket encryption”}, n)$
   (b) Decrypt the next few encrypted tickets with key K_T, until we have ten tickets left unencrypted.

Consider an attacker who gets access to the stored data at some point. He has only ten tickets available. Assuming the KDF is secure, he cannot decrypt any other tickets without access to K_P, which he can get only by successfully using a ticket.

This technique can be used with our basic protocol or with any of our variants, described below.

6.2 Adding Offline Access

It is possible to add a second offline mode of access to the payload key. This may be a practical requirement in many cases, where a user needs to have access even when internet access is not available. However, this represents an explicit tradeoff between security and usability–the number of tickets no longer provides a limit on how many passwords may be guessed by an attacker who compromises the user's device.

If offline access is added, the first question is: how much computation should be required to unlock the payload key offline? We propose the following steps for adding offline access, if this is necessary, making use of the "pepper" idea of Abadi et al. [25]:

1. Determine the largest acceptable amount of computing time on the device that would be acceptable to get offline access. Let this parameter be W. For example, we might require ten minutes' continuous computing on the user's device in order to unlock the offline access. (Note that in many cases, the constraint may be battery life rather than time.)
2. Determine an acceptable time for the initial generation of the offline access information, I, such that $I \times 2^q = W$ for some integer q. For example, I might be 15 seconds on the user's device.
3. Calculate $q \leftarrow \lg(W/I)$.
4. Generate a random salt $S^* \leftarrow_{\$} \{0,1\}^n$ for offline access.
5. Compute $D \leftarrow \mathtt{PH}(S^*, P)$, using parameters for the password hash that require I seconds to compute.
6. Store $Z^* \leftarrow \mathtt{VE}(D, K_P)$
7. Set the low q bits of S^* to zeros.
8. Forget D.

If this is done, we strongly suggest using a memory-hard function for PH, with parameters set to the largest memory requirements that can be reasonably accommodated on the user device–this will make the offline attack more expensive and difficult, and may prevent the attacker using commodity graphics cards to parallelize the attack. Throwing away a few bits of the salt (following Abadi et al.) allows us to only do the work necessary to compute *one* instance of PH during setup, while still requiring an offline user (or attacker) to compute 2^q instances of PH.

6.2.1 Analysis

Consider a user who provides offline access requiring W work, alongside a TPMS scheme for online access. If the user's device is compromised, the attacker is no longer limited to t password guesses–instead, he first makes t password guesses "for free," and then does W work per additional password guess.

Providing offline access throws away one of the major usability advantages available with ticket-mediated password strengthening: the ability of a user to choose a relatively low-entropy password safely. A random dictionary word or six-digit number provides substantial security against an attacker who has only ten guesses. Further, most users can probably understand what kind of passwords are necessary to withstand an attacker who is limited to ten guesses; few can properly estimate whether their password will survive an offline attack given the attacker's budget and the value of W.

The advantage of using a TMPS scheme in this situation is that it allows the work per offline guess to be set to some extremely high value, hopefully making the offline guessing attack too expensive for an attacker in practice, while the user can still get access her data with very little delay as long as she has internet access.

This technique can be used with our basic protocol, or with any of the variants described below. (Though it would not make much sense for the Offline Variant with HSM.)

6.3 An Offline Variant with HSM

Consider the situation where a user has a trusted computing environment or trusted hardware security module (HSM). We define an offline variant of ticket granting and unlocking protocol which uses an HSM and does not need any external server. However, we emphasize that this variant is secure only if the HSM is secure–an attacker who can extract the secret from or reload past states into the HSM can recover the K_P with a simple password search.

6.3.1 Starting Assumptions

We assume that the HSM can be loaded up with a secret value, B, which can not be released from the HSM afterward. We further assume that the HSM supports one-time use of the value B which is updated at each interface as described in Algorithm 5. Note that HSM_Step must be an atomic operation–if any value of D is returned, then B *must* be updated.

We also assume that the user can load a new value of B into the HSM at any time which overwrites the previous existing value.

Algorithm 5 Access Secret and Update HSM Internal State

```
function HSM_STEP($C$)
    $D \leftarrow \mathtt{HMAC}(B, C)$
    $B \leftarrow \mathtt{HASH}(B)$
    return($D$)
```

6.3.2 HSM REQUEST Protocol

In order to generate t tickets, the user first chooses a password, P and generates a random payload key K_P, and then follows the steps listed in Algorithm 6.

Algorithm 6 Create Tickets for the HSM Protocol

```
function HSM_REQUEST($P, K_P, t$)
    $S \leftarrow \{0,1\}^n$
    $B \leftarrow \{0,1\}^n$
    $C \leftarrow \mathtt{PH}(S, P)$
    Load $B$ into the HSM as the new secret value.
    for $i \leftarrow 1 \dots t$ do
        $D_i \leftarrow \mathtt{HMAC}(B, C)$
        $Z_i \leftarrow \mathtt{VE}(D_i, K_P)$
        $B \leftarrow \mathtt{HASH}(B)$
    Forget $D_{1,2,\dots,t}$, $C$, $B$
    return($S, Z_{1,2,\dots,t}$)
```

The protocol uses a fixed random salt S for generating all t tickets. As we do not need privacy from the HSM, reusing the salt and getting same C is acceptable. Similarly, there is no need for public key encryptions or signatures. Guessing the password is equally difficult as the password-derived value C is never stored. The value B is updated after each computation of Z_i, resulting t-different D_i's. The t-tickets $\{Z_1, Z_2, \dots, Z_t\}$ consist *only* of the encryptions of K_P under different keys D_i. As a result, the user storage as single S and $Z_{1,2,\dots,t}$.

Note that the same protocol can be used to generate new tickets when the old ones are running out. In that case, the user simply runs the HSM_UNLOCK algorithm (Algorithm 7) to recover K_P, and then runs the HSM_REQUEST algorithm (Algorithm 6) with P and K_P to get more tickets for the same password and payload key.

6.3.3 HSM UNLOCK Protocol

The process of unlocking the tickets is straightforward; however, it needs sequential run of the protocol starting from ticket number 1 to t and hence, requires a strong synchronization between the HSM and the user. Specifically, the synchronization ensures the computation of the correct value of B and finally the K_P when correct password is provided. The protocol as shown in Algorithm 7 starts with accepting a password $\hat{P}$ from the user, which is used to derive a challenge value $\hat{C}$. The HMAC of this challenge value along with the current value of B inside the HSM is computed by the HSM and returned to the user as $\hat{D}$, and then B is again updated by the HSM. These computations inside the HSM follows the steps of Algorithm 5. Finally, the correctness of $\hat{D}$ is verified by analyzing the output K obtained at Step 4 of the Algorithm 7. The correct value of $\hat{D}$ implies K is the desired K_P.

Algorithm 7 Use an HSM Ticket to Unlock K_P

```
1: function HSM_UNLOCK($\hat{P}, S, Z$)
2:     $\hat{C} \leftarrow \mathtt{PH}(S, \hat{P})$
3:     $\hat{D} \leftarrow$ HSM_Step($\hat{C}$)
4:     $K \leftarrow \mathtt{DV}(\hat{D}, Z_i)$
5:     if $K = \perp$ then
6:         return($\perp$)
7:     else
8:         return($K$)
```

The user must delete all old values of Z, in order to ensure that he can always determine which ticket is to be used next.

It is possible that some software error will lead to the HSM and user software getting out-of-synch. The best strategy for handling this is to attempt to unlock the payload key using the password and the final ticket, and to keep trying until the payload key is unlocked.

6.3.4 Analysis

Note that this scheme is not covered by our security proofs. Here, we provide some arguments for the security of the scheme.

Consider the situation where the user has produced t tickets, and then her laptop was stolen by an attacker who cannot violate the security of the HSM. Informally, what can we say about the attacker's chances of learning K_P?

The attacker needs to guess the correct value of $C = \mathtt{PH}(S, P)$. Each guess of the password leads to a guess of C.

The user has tickets corresponding to the next t values of B that will be used by the HSM. For $i = 1, 2, \dots, t$, a ticket Z_i is used as follows:

$$\begin{aligned} D_i &= \texttt{HMAC}(B_i, C) \\ Z_i &= \texttt{VE}(D_i, K_P) \\ B_{i+1} &= \texttt{HASH}(B_i) \end{aligned}$$

The HSM will only carry out one computation with each value of B_i–this can be used to derive the decryption key D_i, but only if the attacker guesses the right value of C. Each value of B_i inside the HSM allows a new guess of C, and each value of Z_i in the attacker's hands allows the guess to be checked.

After t guesses, the HSM has a value of B for which the attacker has no corresponding values of Z. At this point, the attacker can learn nothing about K_P from interacting with the HSM. Since he also cannot break the encryption, this imposes a hard limit–the attacker gets only t guesses of the password.

In this setting, we have no privacy concerns w.r.t. the HSM. However, it's worth noting that the HSM never sees the password or any value it could use to check a password guess. (Though if the HSM was also used to generate the salt S, this would no longer be true.).

We assume that the HSM is able to securely keep B secret. Along with whatever tamper-resistance features are incorporated into the HSM, since B is updated each time it is used, side-channel attacks are very unlikely to succeed.

6.4 Different Ways to Authorize Tickets

In the basic protocol, we assume that the server issues blind signatures to allow the server to limit how much assistance it is required to provide. (That is, the server's owner presumably wants it to only provide TMPS services to users who have paid for them, or to users who are somehow affiliated with the entity running the server.). A blind signature works well to protect the user's privacy, but makes strong demands on the signature scheme used. In particular, most proposed post-quantum signature schemes have no known blind signature defined. Below, we discuss alternative ways for the server to limit access to its services *without* the need for a blind signature.

Our possible approaches fall into two broad categories:

1. Offline–the REQUEST operation is done without any interaction with the server or any other party.
2. Online–the REQUEST protocol is almost unchanged, but some other operation is substituted for the blind signature protocol.

Note that the security proof on our basic protocol doesn't cover these variants, though we believe it could be modified to cover them without too much difficulty. For each variant, we provide a short sketch of why we believe the variant is secure. Also note that we still assume that the public key encryption used below is CCA2–specifically, it must not be possible to modify a ciphertext without changing the plaintext.

6.4.1 Third Party Signer (Online)

A very lightweight (but imperfect) technique for ensuring user privacy from the server is simply to separate the authorization of getting a ticket from the unlocking of tickets. Suppose we have two trusted parties: the Bank and the Server. The Bank authorizes tickets, and can recognize a ticket, but is never shown tickets by the Server; the Server unlocks tickets but can't recognize them. In this protocol, we need an ordinary signature, nothing more.

The REQUEST protocol works as follows. Note that this is almost the same protocol as with the blind signatures, except it is done with the Banker instead of the Server.

```
Protocol: ThirdParty_REQUEST$(P, K_P, t)$:
User                                   Banker
for $i = 1 \ldots t$
    $B \leftarrow_{\$} \{0,1\}^n$
    $E \leftarrow \texttt{ENC}(PK_S, B)$
                          $\xrightarrow{E}$
                                           $F \leftarrow \text{sign}(SK_B, E)$
                          $\xleftarrow{F}$
    $S \leftarrow_{\$} \{0,1\}^n$
    $C \leftarrow \texttt{PH}(S, P)$
    $D \leftarrow \texttt{HMAC}(B, C)$
    $Z \leftarrow \texttt{VE}(D, K_P)$
    Forget $B, C, D, E^*, \sigma^*$
    $T_i \leftarrow (S, E, F, Z)$
endfor
return$(T_{1,2,\ldots,t})$
```

The unlocking protocol is exactly the same except for the public key used to verify the signature.

Security This scheme is almost identical to the basic protocol–the only difference is that the REQUEST protocol is run with a different server, and no blind signature is used. The user's privacy from the server is ensured by separation of information–the banker knows the signatures it issued to the user, but doesn't share that information with the server. An attacker who compromises the user's device has exactly the same probability of success in this scheme as in the main protocol.

6.4.2 Group Signatures (Offline)

If we want to use TMPS with a signature scheme which doesn't allow blind signatures, but allows group or ring signatures, then a small variation of the protocol can be done. We assume here that the user has a group public key PK and a personal private key SK_U for the group signature scheme. We also assume that the server knows PK.

Algorithm 8 Use Group Signature to Create Tickets

```
function GROUP_REQUEST($P, K_P, t$)
    for $i \leftarrow 1 \dots t$ do
        $B \leftarrow_{\$} \{0,1\}^n$
        $E \leftarrow \mathtt{ENC}(PK_S, B)$
        $F \leftarrow \mathrm{GroupSign}(SK_U, E)$
        $S \leftarrow_{\$} \{0,1\}^n$
        $C \leftarrow \mathtt{PH}(S, P)$
        $D \leftarrow \mathtt{HMAC}(B, C)$
        $Z \leftarrow \mathtt{VE}(D_i, K_P)$
        Forget $D_{1,2,\dots,N}$, $C$, $B$
        $T_i \leftarrow (S, E, F, Z)$
    return($T_{1,2,\dots,t}$)
```

The corresponding UNLOCK protocol is almost unchanged–the server simply verifies that F signs E using the group public key PK rather than its own signature public key PK'_S.

Security Consider an attacker who compromises the user's device, and thus learns his tickets and his signing key. The attacks possible to him are the same as in the basic protocol–he can request new tickets, but without knowing K_P or P, these will give him no help in learning the correct value of P or K_P. Despite knowing the signing key, he cannot alter tickets to give himself more guesses, because the public key encryption used is CCA2.

The server can't learn which user is UNLOCKing her key at any given time, because the group signature tells the server only that she is a member of the group.

6.4.3 Proof of Work (Offline)

If the server's main concern is having its resources wasted rather than being paid for its services, a simple alternative is to require a proof of work for each new ticket. Let's add two new functions:

$y \leftarrow \mathtt{MakePOW(x,W)}$ does approximately W work to create a proof of work, y, associated with input value x.

$\mathtt{CheckPOW(x,y,W)}$ returns 1 if the proof of work is valid, and 0 otherwise.

With these two, we can define an entirely offline proof-of-work version of our protocol, where W is assumed to be a known systemwide parameter.

Algorithm 9 Use Proof of Work to Create Tickets

```
function POW_REQUEST($P, K_P, t$)
    for $i \leftarrow 1 \dots t$ do
        $B \leftarrow_{\$} \{0,1\}^n$
        $E \leftarrow \mathtt{ENC}(PK_S, B)$
        $F \leftarrow \mathtt{MakePOW}(E, W)$
        $S \leftarrow_{\$} \{0,1\}^n$
        $C \leftarrow \mathtt{PH}(S, P)$
        $D \leftarrow \mathtt{HMAC}(B, C)$
        $Z \leftarrow \mathtt{VE}(D_i, K_P)$
        Forget $D_{1,2,\dots,N}$, $C$, $B$
        $T_i \leftarrow (S, E, F, Z)$
    return($T_{1,2,\dots,t}$)
```

Protocol: POW_UNLOCK($S, E, F, Z, \hat{P}$):

```
User                                   Server
$\hat{C} \leftarrow \mathtt{PH}(S, \hat{P})$
                $\xrightarrow{E, F, \hat{C}}$
                                       IF
                                           $E$ fresh AND
                                           $\mathtt{CheckPOW}(F, E, W)$
                                       THEN
                                           $B \leftarrow \mathtt{DEC}(SK_S, E)$
                                           $D \leftarrow \mathtt{HMAC}(B, \hat{C})$
                                       ELSE
                                           $D \leftarrow \perp$
                $\xleftarrow{D}$
$K_P \leftarrow \mathtt{DV}(D, Z)$
return($K_P$)
```

Security The use of the proof of work eliminates any information about the user in the ticket, and in fact, makes the REQUEST protocol non-interactive.

An attacker who compromises the user's device can make up additional tickets, but without knowing K_P or P, these do not help him learn the correct values of P or K_P. Again, the attacker cannot alter the value of E, because the public key encryption is CCA2 secure.

7 Implementation

The TMPS protocol requires the use of a public key encryption scheme (e.g. RSA or El Gamal), a hash function (e.g. SHA2, SHA3, or Blake2) and an HMAC computation, a password-hashing scheme (e.g. PBKDF2, scrypt, or Argon2), a blind signature protocol (e.g. RSA or El Gamal). The protocol permits a great deal of flexibility in choice of underlying cryptographic primitives. Notably, all of these (except possibly the last) can be accomplished using existing quantum-resistant algorithms.

We implemented our protocol in Python [3], using the Cryptography module, which provides a Python frontend for OpenSSL calls. The protocol allows a choice of underlying primitives; we used RSA with 3072-bit moduli for signing and public key encryption, along with SHA256 for hashing, and PBKDF2_HMAC_SHA2 for password-hashing.

All measurements were performed on a Macbook Pro (3.5 GHz Intel Core i7).While this is not an optimized implementation, it allows us to obtain concrete performance numbers.

Requesting 1000 tickets took a total of 83.9 seconds, of which 76.3 was taken up by the server. This was mostly the work of doing an RSA blind signature–the library we used didn't have a blind signature call, so we implemented one ourselves–this is much slower than the library call. In our implementation, each ticket took about 0.0763 seconds to produce on the server side, and about 0.0076 seconds on the user side. The blind signature done on the server side consists of a modular exponentiation with the exponent d, and should take no more time than a normal RSA signature. A better estimate for an optimized Python implementation's server work would be about 1.6 seconds for 1000 tickets, or about 0.0016 seconds per requested ticket. Unlocking 1000 tickets took a total of 6.7 seconds, of which 1.8 seconds was taken up by the server. Thus, the server required 0.0018 seconds to assist in the unlocking of a ticket, while the user needed about 0.0049 seconds. We provide these performance numbers to demonstrate that the protocol is practical–even with an unoptimized Python implementation.

8 Conclusion and Open Questions

In this paper, we have proposed a new mechanism for strengthening password-based key derivations, called TMPS (Ticket-Mediated Password Strengthening). We have also proposed a set of protocols that implements a TMPS scheme, proved its security in the UC model, and provided a number of variant schemes which allow for different implementation constraints and tradeoffs.

There are several questions left open by this research.

First, whether it is possible to construct TMPS schemes which provide privacy for the user, allow the server to restrict access to only authorized users, and do not need blind or group signatures. The variant protocols in Section 6 that avoided these primitives imposed other requirements–either a willingness to trust a third party with user privacy, a willingness to provide the service to all comers, or the use of group signatures.

Second, investigating other settings where one can use tickets bound to a computation to obtain a novel functionality. For example, it may be possible to use this kind of mechanism to limit accesses to a local encrypted database, or computations of a key derivation function.

Third, whether there are more elaborate restrictions that can be imposed on these tickets, without losing the users' privacy. For example, is it be possible to rate-limit UNLOCK requests from a given user without revealing which user was using the scheme?

Finally, incorporating a mechanism for key rollover, for situations where a server's key may have been compromised, would be a useful addition to the scheme. At present, our solution would be to REQUEST a large set of replacement tickets using the server's new public key.

References

[1] AGRAWAL, S., MIAO, P., MOHASSEL, P., AND MUKHERJEE, P. PASTA: password-based threshold authentication. In *ACM Conference on Computer and Communications Security* (2018), ACM, pp. 2042–2059.

[2] AKHAWE, D. How dropbox securely stores your passwords. https://blogs.dropbox.com/tech/2016/09/how-dropbox-securely-stores-your-passwords/, 2016. Online; accessed 4 January 2019.

[3] We will make its source code available on a public-facing git repository

[3] Bellare, M., Keelveedhi, S., and Ristenpart, T. Dupless: Server-aided encryption for deduplicated storage. *IACR Cryptology ePrint Archive 2013* (2013), 429.

[4] Bellare, M., Micciancio, D., and Warinschi, B. Foundations of group signatures: Formal definitions, simplified requirements, and a construction based on general assumptions. In *EUROCRYPT 2003* (May 2003), E. Biham, Ed., vol. 2656 of *LNCS*, Springer, Heidelberg, pp. 614–629.

[5] Bellare, M., Pointcheval, D., and Rogaway, P. Authenticated key exchange secure against dictionary attacks. In *Proceedings of the 19th International Conference on Theory and Application of Cryptographic Techniques* (Berlin, Heidelberg, 2000), EUROCRYPT'00, Springer-Verlag, pp. 139–155.

[6] Bellovin, S. M., and Merritt, M. Encrypted key exchange: Password-based protocols secure against dictionary attacks. In *IEEE SYMPOSIUM ON RESEARCH IN SECURITY AND PRIVACY* (1992), pp. 72–84.

[7] Camenisch, J., Drijvers, M., Gagliardoni, T., Lehmann, A., and Neven, G. The wonderful world of global random oracles. In *EUROCRYPT 2018, Part I* (Apr. / May 2018), J. B. Nielsen and V. Rijmen, Eds., vol. 10820 of *LNCS*, Springer, Heidelberg, pp. 280–312.

[8] Camenisch, J., Enderlein, R. R., and Neven, G. Two-server password-authenticated secret sharing UC-secure against transient corruptions. Cryptology ePrint Archive, Report 2015/006, 2015. http://eprint.iacr.org/2015/006.

[9] Canetti, R. Security and composition of multiparty cryptographic protocols. *Journal of Cryptology 13*, 1 (Jan. 2000), 143–202.

[10] Canetti, R., Damgård, I., Dziembowski, S., Ishai, Y., and Malkin, T. On adaptive vs. nonadaptive security of multiparty protocols. In *EUROCRYPT 2001* (May 2001), B. Pfitzmann, Ed., vol. 2045 of *LNCS*, Springer, Heidelberg, pp. 262–279.

[11] Canetti, R., Feige, U., Goldreich, O., and Naor, M. Adaptively secure multi-party computation. In *28th ACM STOC* (May 1996), ACM Press, pp. 639–648.

[12] Everspaugh, A., Chaterjee, R., Scott, S., Juels, A., and Ristenpart, T. The pythia PRF service. In *24th USENIX Security Symposium (USENIX Security 15)* (Washington, D.C., 2015), USENIX Association, pp. 547–562.

[13] Ford, W., and Jr., B. S. K. Server-assisted generation of a strong secret from a password. In *9th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2000), 4-16 June 2000, Gaithersburg, MD, USA* (2000), IEEE Computer Society, pp. 176–180.

[14] Hohenberger, S., Lewko, A., and Waters, B. Detecting dangerous queries: A new approach for chosen ciphertext security. Cryptology ePrint Archive, Report 2012/006, 2012. http://eprint.iacr.org/2012/006.

[15] Jarecki, S., Kiayias, A., Krawczyk, H., and Xu, J. TOPPSS: Cost-minimal password-protected secret sharing based on threshold OPRF. In *ACNS 17* (July 2017), D. Gollmann, A. Miyaji, and H. Kikuchi, Eds., vol. 10355 of *LNCS*, Springer, Heidelberg, pp. 39–58.

[16] Jarecki, S., Krawczyk, H., and Xu, J. OPAQUE: an asymmetric PAKE protocol secure against pre-computation attacks. In *EUROCRYPT (3)* (2018), vol. 10822 of *Lecture Notes in Computer Science*, Springer, pp. 456–486.

[17] Katz, J., and Lindell, Y. *Introduction to Modern Cryptography, Second Edition*. CRC Press, 2014.

[18] Kelsey, J., Schneier, B., Hall, C., and Wagner, D. Secure applications of low-entropy keys. In *ISW'97* (Sept. 1998), E. Okamoto, G. I. Davida, and M. Mambo, Eds., vol. 1396 of *LNCS*, Springer, Heidelberg, pp. 121–134.

[19] Krawiecka, K., Kurnikov, A., Paverd, A., Mannan, M., and Asokan, N. Safekeeper: Protecting web passwords using trusted execution environments. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018* (2018), P. Champin, F. L. Gandon, M. Lalmas, and P. G. Ipeirotis, Eds., ACM, pp. 349–358.

[20] Krawiecka, K., Paverd, A., and Asokan, N. Protecting password databases using trusted

hardware. In *Proceedings of the 1st Workshop on System Software for Trusted Execution, SysTEX@Middleware 2016, Trento, Italy, December 12, 2016* (2016), ACM, pp. 9:1–9:6.

[21] LAI, R. W. F., EGGER, C., REINERT, M., CHOW, S. S. M., MAFFEI, M., AND SCHRÖDER, D. Simple password-hardened encryption services. In *27th USENIX Security Symposium (USENIX Security 18)* (Baltimore, MD, 2018), USENIX Association, pp. 1405–1421.

[22] LAI, R. W. F., EGGER, C., SCHRÖDER, D., AND CHOW, S. S. M. Phoenix: Rebirth of a cryptographic password-hardening service. In *26th USENIX Security Symposium (USENIX Security 17)* (Vancouver, BC, 2017), USENIX Association, pp. 899–916.

[23] LINDELL, A. Y. Adaptively secure two-party computation with erasures. In *CT-RSA 2009* (Apr. 2009), M. Fischlin, Ed., vol. 5473 of *LNCS*, Springer, Heidelberg, pp. 117–132.

[24] MANI, A. Life of a password. Real World Crypto 2015, 2015. https://rwc.iacr.org/2015/Slides/RWC-2015-Amani.pdf.

[25] MARTIN ABADI, T. M. A. L., AND NEEDHAM, R. Strengthening passwords. Technical report, 1997. https://users.soe.ucsc.edu/~abadi/Papers/pwd-revised/pwd-revised.html.

[26] MUFFETT, A. Facebook: Password hashing & authentication. Presentation at Passwords 2014 Conference, NTNU, 2014. https://video.adm.ntnu.no/pres/54b660049af94.

[27] MUFFETT, A. Life of a password. Presentation at Real World Crypto 2015, 2015.

[28] SCHNEIDER, J., FLEISCHHACKER, N., SCHRÖDER, D., AND BACKES, M. Efficient cryptographic password hardening services from partially oblivious commitments. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, Austria, October 24-28, 2016* (2016), E. R. Weippl, S. Katzenbeisser, C. Kruegel, A. C. Myers, and S. Halevi, Eds., ACM, pp. 1192–1203.

[29] SCHRÖDER, D., AND UNRUH, D. Security of blind signatures revisited. In *PKC 2012* (May 2012), M. Fischlin, J. Buchmann, and M. Manulis, Eds., vol. 7293 of *LNCS*, Springer, Heidelberg, pp. 662–679.

[30] SÖNMEZ TURAN, M., BARKER, E. B., BURR, W. E., AND CHEN, L. SP 800-132. recommendation for password-based key derivation: Part 1: Storage applications. Tech. rep., National Institute of Standards & Technology, Gaithersburg, MD, United States, 2010.

[31] WU, T. The SRP authentication and key exchange system. *RFC 2945* (2000), 1–8.

A Definitions

In this section, we mention the key definitions used in the security analysis of our protocol to facilitate better understanding. Our exposition closely follows [4, 14, 17, 29].

Definition A.1. [Encryption System] An encryption system can be defined as a tuple of probabilistic polynomial-time algorithms $\Pi_{\texttt{ENC}}(\texttt{GEN}, \texttt{ENC}, \texttt{DEC})$ such that:

1. The key-generation algorithm $\texttt{GEN}$ takes as input the security parameter 1^n and outputs a key K.
2. The encryption algorithm $\texttt{ENC}$ takes as input a key K and a plaintext message $M \in \{0,1\}^*$, and outputs a ciphertext C where $C \leftarrow \texttt{ENC}_K(M)$.
3. The decryption algorithm $\texttt{DEC}$ takes as input a key and a ciphertext, and outputs a message. We assume without loss of generality that the decryption algorithm corresponding $\texttt{ENC}_K$ is $\texttt{DEC}_K$ such that $M = \texttt{DEC}_K(C)$ and for every n, every key K output by $\texttt{GEN}(1^n)$, and every $M \in \{0,1\}^*$, it holds that $\texttt{DEC}_K(\texttt{ENC}_K(M)) = M$.

The Chosen-Ciphertext Attack (CCA) security experiment $PrivK^{cca}_{\mathcal{A},\Pi_{\texttt{ENC}}}(n)$**:** Consider the following experiment for an encryption system $\Pi_{\texttt{ENC}} = (\texttt{GEN}, \texttt{ENC}, \texttt{DEC})$, adversary $\mathcal{A}$, and value n for the security parameter.

1. A random key K is generated by running $\texttt{GEN}(1^n)$.
2. The adversary $\mathcal{A}$ is given input 1^n and oracle access to $\texttt{ENC}_K(\cdot)$ and $\texttt{DEC}_K(\cdot)$. It outputs a pair of messages M_0, M_1 of the same length.
3. A random bit $b \leftarrow \{0,1\}$ is chosen, and then a ciphertext $C \leftarrow \texttt{ENC}_K(M_b)$ is computed and given to $\mathcal{A}$. We call C the challenge ciphertext.
4. The adversary $\mathcal{A}$ continues to have oracle access to $\texttt{ENC}_K(\cdot)$ and $\texttt{DEC}_K(\cdot)$, but is not allowed to query the latter on the challenge ciphertext itself. Eventually, $\mathcal{A}$ outputs a bit b'

5. The output of the experiment is defined to be 1 if $b' = b$, and 0 otherwise.

Definition A.2. [CCA Security] An encryption system $\Pi_{\mathtt{ENC}}$ has indistinguishable encryptions under a chosen-ciphertext attack (or is CCA-secure) if for all probabilistic polynomial-time adversaries $\mathcal{A}$ there exists a negligible function negl such that:

$$Pr[PrivK^{cca}_{\mathcal{A},\Pi_{\mathtt{ENC}}}(n) = 1] \leq \frac{1}{2} + negl(n),$$

where the probability is taken over all random coins used in the experiment.

Other variants of the CCA Security definition are defined below.

Definition A.3. [Chosen Plaintext Attack (CPA) Security] Similar to the security experiment of CCA except that the Adversary $\mathcal{A}$ is not given access to decryption oracle at step 2 and step 4.

Definition A.4. [Non-adaptive CCA or CCA1 Security] Similar to the security experiment of CCA except that the Adversary $\mathcal{A}$ is not given access to decryption oracle at step 4.

Definition A.5. [Adaptive CCA or CCA2 Security] Similar to the security experiment of CCA where the Adversary $\mathcal{A}$ is allowed to perform a polynomially bounded number of encryptions, decryptions or other calculations over inputs of its choice except on the challenge ciphertext.

Definition A.6. [Signature Scheme]: A signature scheme is a tuple of probabilistic polynomial-time algorithms $\Pi_{SIG}(\mathtt{GEN}, \mathtt{SIGN}, \mathtt{VERIFY})$ such that:

1. The key-generation algorithm GEN takes as input a security parameter 1^n and outputs a pair of keys (PK, SK). These are called the public key and the private key, respectively.
2. The signing algorithm SIGN takes as input a private key SK and a message M from some underlying message space. It outputs a signature F represented as $F \leftarrow \mathtt{SIGN}_{SK}(M)$.
3. The deterministic verification algorithm VERIFY takes as input a public key PK, a message M, and a signature F. It outputs a bit b represented as $b = \mathtt{VERIFY}_{PK}(M, F)$ where $b = 1$ means valid and $b = 0$ means invalid.

We require that for every n, every (PK, SK) output by $\mathtt{GEN}(1^n)$, and every message M in the appropriate underlying plaintext space, it holds that

$$\mathtt{VERIFY}_{PK}(M, \mathtt{SIGN}_{SK}(M)) = 1.$$

We say F is a valid signature on a message M if $\mathtt{VERIFY}_{PK}(M, F) = 1$.

Definition A.7. [Blind Signature] A 2-move blind signature scheme is an interactive signature scheme with signer $\mathcal{S}$ and user $\mathcal{U}$ and can be defined as a tuple of probabilistic polynomial-time algorithms $\Pi_{BSIG} = (\mathtt{GEN}, \mathtt{BLIND}, \mathtt{UBLIND}, \mathtt{SIGN}, \mathtt{BVERIFY})$ such that:

1. The key-generation algorithm Gen takes as input a security parameter 1^n and outputs a pair of keys (PK, SK). These are called the public key and the private key, respectively.
2. Signature Issuing. The parties execute the following protocol, denoted $\langle \mathcal{U}(PK, M), \mathcal{S}(SK) \rangle$:
   (a) $M^* \leftarrow \mathtt{BLIND}(M)$: The user blinds the message M to obtain M^* and sends to the signer.
   (b) $F^* \leftarrow \mathtt{SIGN}_{SK}(M^*)$: The signer outputs a signature F^* on input of message M^* and private key SK and sends to the user.
   (c) $F \leftarrow \mathtt{UBLIND}(F^*)$: The user unblinds the signature F^* to obtain F. Note that the user inputs additional private state to the UBLIND algorithm, which we leave implicit.
3. The deterministic verification algorithm BVERIFY takes as input a public key PK, a message M, and a signature F. It outputs a bit b where $b = 1$ means valid and $b = 0$ means invalid.

We require that for every n, every (PK, SK) output by $\mathtt{GEN}(1^n)$, and every message $M \in \{0, 1\}^n$ and any F output by $\mathcal{U}$ in the joint execution of $\langle \mathcal{U}(PK, M), \mathcal{S}(SK) \rangle$, it holds that

$$\mathtt{BVERIFY}_{PK}(M, F) = 1.$$

The security of blind signature schemes requires two properties, namely unforgeability and blindness.

Definition A.8. [Unforgeability] A 2-move blind signature scheme $\Pi_{BSIG} = (\mathtt{GEN}, \mathtt{BLIND}, \mathtt{UBLIND}, \mathtt{SIGN}, \mathtt{BVERIFY})$ is called unforgeable if for any efficient algorithm $\mathcal{A}$ the probability that experiment $\mathrm{Unforge}^{\Pi_{BSIG}}_{\mathcal{A}}(n)$ evaluates to 1 is negligible (as a function of n) where

Experiment $\mathbf{Forge}^{\mathcal{A}}_{\Pi_{BSIG}}$

1. $(SK, PK) \leftarrow \texttt{GEN}(1^n)$
2. $((M_1,F_1),\cdots,(M_{k+1},F_{k+1})) \leftarrow \mathcal{A}^{\langle\cdot,\mathcal{S}(SK)\rangle^{\infty}}(PK)$ Return 1 iff
   (a) $M_i \neq M_j$ for $1 \leq i < j \leq k+1$ and
   (b) $\texttt{BVERIFY}_{PK}(M_i, F_i) = 1$ for all $i = 1, 2, \cdots, k+1$, and
   (c) at most k interactions with $\langle\cdot,\mathcal{S}(SK)\rangle^{\infty}$ were completed.

Definition A.9. [Blindness] A 2-move blind signature scheme $\Pi_{BSIG} = (\texttt{GEN}, \texttt{BLIND}, \texttt{UBLIND}, \texttt{SIGN}, \texttt{BVERIFY})$ is called blind if for any efficient algorithm $\mathcal{A}$ the probability that experiment $Blind^{\Pi_{BSIG}}_{\texttt{BSIGN}^*}(n)$ evaluates to 1 is negligibly close to $\frac{1}{2}$ where

Experiment Blind$^{\Pi_{BSIG}}_{\texttt{BSIGN}^*}$

1. $(PK, M_0, M_1, st_{find}) \leftarrow \mathcal{A}(find, 1^n)$
2. $b \leftarrow \{0,1\}$
3. $st_{issue} \leftarrow \mathcal{A}^{\langle\mathcal{U}(PK,M_b),\cdot\rangle^1,\langle\mathcal{U}(PK,M_{1-b}),\cdot\rangle^1}(issue, st_{find})$ and let F_b, F_{1-b} denote the (possibly undefined) local outputs of $\mathcal{U}(PK, M_b)$ resp. $\mathcal{U}(PK, M_{1-b})$
4. set $(F_0, F_1) = (\perp, \perp)$ if $F_0 = \perp$ or $F_1 = \perp$
5. $b^* = \mathcal{A}(guess, F_0, F_1, st_{issue})$
6. return 1 iff $b = b^*$.

Definition A.10. [Group Signature] A group signature scheme $\Pi_{GSIG} = (GK_g, \texttt{GSIGN}, \texttt{GVERIFY}, \texttt{OPEN})$ consists of four polynomial-time algorithms:

1. The randomized group key generation algorithm GK_g takes input a security parameter 1^n and 1^m where $m \in \mathbb{N}$ is the group size and outputs a tuple $(gPK, gmSK, gSK)$, where gPK is the group public key, $gmSK$ is the group manager's secret key, and gSK is an n-vector of keys with $gSK[i]$ being a secret signing key for player $i \in [m]$.
2. The randomized group signing algorithm GSIGN takes as input a secret signing key $gSK[i]$ and a message M to return a signature of M under $gSK[i]$ $i \in [m]$.
3. The deterministic group signature verification algorithm GVERIFY takes as input the group public key gPK, a message M, and a candidate signature F for M to return either 1 or 0.
4. The deterministic opening algorithm OPEN takes as input the group manager secret key $gmSK$, a message M, and a signature F of M to return an identity i or the symbol $\perp$ to indicate failure.

Correctness: The scheme must satisfy the following correctness requirement. For all $n, m \in \mathbb{N}$, all $(gPK, gmSK, gSK) \in [GK_g(1^n, 1^m)]$, all $i \in [n]$ and all $M \in \{0,1\}^*$

$$\texttt{GVERIFY}(gPK, M, \texttt{GSIGN}(gSK[i], M)) = 1 \text{ and}$$

$$\texttt{OPEN}(gmSK, M, \texttt{GSIGN}(gSK[i], M)) = i$$

Definitions of security in the Universal Composability (UC) framework. We refer to previous work [9, 10, 23] for definitions of UC secure computation in the adaptive-corruption setting.

B Game-Based Security Definitions and Proofs

Theorem 5.1 ensures that the basic protocol behaves like the ideal functionality, but does not tell us exactly what security properties the ideal functionality provides. In this section, we consider some game-based security definitions, and show that the ideal functionality makes it easy to prove a bound on an attacker's probability of winning these games. Combined with Theorem 5.1, we thus prove that no attacker interacting with our basic protocol can win these games with probability more than negligibly higher than these bounds.

B.1 User Compromise: Stealing Tickets

The most practically important attack to consider involves the compromise of a user's data. For example, Bob steals Alice's laptop with an encrypted hard drive; he knows all her tickets, and can impersonate her to the server, but he doesn't know Alice's password. The security goal of our scheme in this case is straightforward: after Bob steals Alice's t tickets, he gets only t guesses for her password. We can define this in terms of the following game:

Security Game: User Compromise

1. The game is parametrized by security parameter n, dictionary size N and number of tickets t, where $t < N$.
2. The challenger generates a list of N distinct passwords, $P_{1,\ldots,N}$.
3. The challenger randomly generates a payload key, $K_P \leftarrow \{0,1\}^n$.

4. The challenger chooses a random $\ell \in \{1,\ldots,N\}$.
5. The challenger honestly runs `ServerSetup` and `RequestTickets` the to generate t tickets $T_{1,\ldots,t}$ using password P_ℓ and payload key K_P.
6. The challenger provides the attacker with the list of N passwords and t tickets.
7. The attacker may send any messages he likes to the server, and may do any computation he likes, up to some very generous limits.
8. The attacker may request new tickets by running `RequestTickets` with the challenger.
9. The attacker must return a guess of K_P. If his guess is correct, he wins; otherwise he loses.

In this game, we explicitly assume that the server being used is *not* compromised.

Consider an attacker allowed to make at most 2^k queries to the server, and at most 2^k trial decryption attempts to `DV`. The protocol meets its security target against this attacker if he wins the game with probability at most $\frac{t}{N} + \mathsf{negl}(n-k)$.

B.1.1 Practical Relevance

This game directly relates to the attack we are most concerned with in this system–the one where Bob learns all of Alice's stored information, and tries to guess her password so he can decrypt all her files. If Alice keeps only t tickets on her computer, this must translate to Bob getting no more than t guesses at her password, in order for our system to be secure.

B.1.2 Proof

We now will prove that an adversary interacting with the ideal functionality and limited to at most 2^k trial decryptions with `DV` has at most a $\frac{t}{N} + 2^{k-n}$ probability of winning this game.

Definition B.1. *Winning Transcript*
An attacker has a *winning transcript* (WT) when the transcript of his interactions with the ideal functionality includes at least one response to an UNLOCK request which contains the correct value of K_P.

A *losing transcript* (LT) is any transcript which is not a winning transcript.

Fact. Given a winning transcript, the attacker has (at most) a probability of one of winning the game.

This is trivially true of any condition, though an attacker with a winning transcript actually has *exactly* a probability of one of winning the game.

We write $Pr[WT]$ to denote the probability of getting a winning transcript, and $Pr[WT|A]$ to denote the probability of adversary A getting a winning transcript.

Definition B.2. *Well-Behaved Adversary*
Consider an adversary given tickets $T_{1,2,\ldots,t}$ and passwords $P_{1,2,\ldots,N}$. A well-behaved adversary (WBA) makes queries to the ideal functionality according to the following rules:

1. Never make a REQUEST query.
2. For each UNLOCK query:
   (a) Use a ticket $T \in T_{1,\ldots,t}$.
   (b) Never use the same ticket in more than one UNLOCK query.
   (c) Use a password $\hat{P} \in P_{1,\ldots,N}$.
   (d) Never use the same password in more than one UNLOCK query.

Lemma B.1. A WBA has a probability of at most $\frac{t}{N}$ of having a winning transcript.

Proof:

1. Every WBA UNLOCK query has a valid ticket. (Definition of WBA (a,b) and Definition of Game (step 5))
2. Every WBA UNLOCK query returns $(1,\perp)$ if its password is incorrect, and $(1,K_P)$ if its password is correct. (Ideal functionality, lines 26 and 29)
3. WBA makes at most t queries. (WBA definition, (a,b))
4. $P[GT] = P[$ correct password appears in one of WBA's UNLOCK queries$]$. (Implication of step 2.)
5. $P[$correct password appears in one of WBA's queries$] \leq \frac{t}{N}$
   (a) WBA makes t queries, each with a different password.
   (b) One password is correct, but adversary has no information about which.

(c) Thus, WBA has at best a $\frac{t}{N}$ chance of including the right password in one of its queries.

6. Thus, a WBA has at most a $\frac{t}{N}$ probability of having a winning transcript. (Previous two steps.)

Lemma B.2. There is no adversary A such that $Pr[WT|A] > Pr[WT|WBA] + 2 \times 2^{k-n}$.

Proof: By showing that violating any of the five conditions of being a WBA can never raise $P[WT]$ by more than a negligible amount.

1. Making a REQUEST call never raises $P[WT]$.
   (a) A WT is a transcript in which an UNLOCK returns K_P.
   (b) Tickets generated by a REQUEST permit an UNLOCK with the same payload key as appeared in the REQUEST.
   (c) The value of K_P is selected randomly from all possible n-bit strings.
   (d) The adversary may make at most 2^k REQUEST calls.
   (e) Thus, the probability of the adversary putting the right value of K_P in one of those REQUEST calls (which will enable an UNLOCK call with the right value of K_P) is no greater than 2^{k-n}.
2. Using a ticket that's not in the valid set of tickets never raises $P[WT]$. (Only tickets $T_{1,2,...,t}$ have payload key K_P.
3. Reusing a ticket never raises $P[WT]$. (A reused ticket always gets $(\perp, \perp)$.
4. Using a password that's not in $P_{1,...,N}$ never raises $P[WT]$. (Only an UNLOCK with P_ℓ will get back $(1, K_P)$, any other password will get $(1, \perp)$.)
5. Reusing a password never raises $P[WT]$. (If the password was used with a valid fresh ticket in a previous UNLOCK call, then future UNLOCK calls with fresh tickets will get the same result. Thus, the probability of K_P appearing in the transcript is never raised.)
6. Thus, $P[WT|\text{non-}WBA] \leq P[WT|WBA] + 2^{k-n}$.

□

Theorem B.3. With the ideal functionality and an uncompromised server, no adversary who can make at most 2^k queries to `DV` or the ideal functionality can win this game with probability higher than $\frac{t}{N} + 2 \times 2^{k-n}$.

Proof:

1. **GT:** No adversary has more than $\frac{t}{N} + 2 \times 2^{k-n}$ probability of getting a winning transcript when interacting with the ideal functionality. (Lemma B.2.)
2. **LT:** Given a losing transcript, an adversary who can make no more than 2^k trial decryptions with `DV` has a probability of at most 2^{k-n} of determining K_P. (Game definition with K_P chosen randomly.)
3. **Union Bound:** The probability of the attacker knowing K_P is thus no higher than $\frac{t}{N} + 3 \times 2^{k-n}$ (Summing GT and LT conditions.)

□

B.2 Server Compromise: Learning the User's Password

Another critical security property of this scheme is that the server must never learn anything about the user's password. We capture this with the following game, in which we assume the server is corrupted:

Security Game: Learn User's Password

1. The game is parametrized by security parameter n, dictionary size N and number of tickets t, where $t < N$.
2. The challenger generates two random passwords, P_1, P_2.
3. The challenger randomly generates a payload key, $K_P \leftarrow \{0,1\}^n$.
4. The attacker is allowed to play the role of the server in the protocol.
5. The challenger honestly runs `ServerSetup` and `RequestTickets` with the attacker playing the role of the server, using password P_1 and payload key K_P. to generate t tickets $T_{1,...,t}$.
6. The challenger runs UNLOCK using password P_1 and each ticket in succession.
7. The challenger generates a random bit b, and sends the attacker $P_b, P_{b\oplus 1}$.
8. The attacker must guess b to win the game.

Kelsey, John M.; Dachman-Soled, Dana; Sonmez Turan, Meltem; Mishra, Sweta. "TMPS: Ticket-Mediated Password Strengthening." Paper presented at The Cryptographer's Track of the RSA Conference (CT-RSA 2020), San Francisco, CA, US. February 24, 2020 - February 28, 2020.

B.2.1 Practical Relevance

If a compromised server can learn anything about the user's password, then it becomes a major security threat–a single server being compromised might lead to the leak of thousands of users' passwords. If there's no attacker who can win this game with probability more than $\frac{1}{2}$, then the server learns nothing at all about the user's password–not even enough to distinguish the correct password from an incorrect one when given both values. An attacker who can't distinguish correct and incorrect passwords also cannot mount a brute-force password search.

B.2.2 Proof

Theorem B.4. No attacker can win the Learn User's Password game when the attacker and challenger are interacting via the ideal functionality with probability higher than $\frac{1}{2}$.

Proof: In the ideal functionality, the server never receives any information about the password P provided by the user. With no information about the correct password, the attacker has no strategy better than a random guess for determining b. □

B.3 Server Compromise: Violating the User's Privacy

The user trusts the server to assist her in key derivation, but may not want the server to be able to determine when she is deriving her key. This game captures a critical privacy property–the server must not be able to determine which user is unlocking her key at any given time.

Security Game: Violate User Privacy

1. The game is parametrized by security parameter n, and number of tickets t.
2. The challenger generates two random passwords, P_1, P_2, and two payload keys K_P1 and K_P2.
3. The attacker is allowed to play the role of the server in the protocol.
4. The challenger honestly runs `ServerSetup`.
5. The challenger honestly runs REQUEST to generate t tickets with password P_1 and payload key K_P1, identifying itself as user 1.
6. The challenger honestly runs REQUEST to generate t tickets with password P_2 and payload key K_P2, identifying itself as user 2.
7. For $i = 1 \dots t-1$:
   (a) The challenger asks the attacker which user should make the next UNLOCK call, and whether he should use the right password or not.
   (b) The challenger makes the UNLOCK call as directed.
8. The challenger generates a random bit b.
9. The challenger runs UNLOCK using password P_b and one of user $b+1$'s tickets.
10. The attacker must guess b to win the game.

B.3.1 Practical Relevance

We want to guarantee that the user retains privacy from the server–she doesn't give the server the power to track each time she decrypts her hard drive. This game captures the user's privacy goal–an attacker who has compromised the server cannot learn which user ran the UNLOCK protocol with him in any given instance, even if he knows which user requested each ticket and has observed many other uses of the same password. Note that this assumes that the server isn't able to simply track the user by IP address–network-level anonymization of the user is outside the scope of our work.

B.3.2 Proof

The proof is trivial: the ideal functionality does not inform the server which user is making an UNLOCK call, so the server dealing with the ideal functionality never learns this information.

C Proof of Theorem 5.1

Before re-stating our theorem, we note that the only random oracle that gets programmed in the proof is H_{VE}.[4]

[4] We note that for UC composition to hold in the programmable random oracle model, one must, in general, assume that an independent random oracle is used for each SID instance. In our case, we essentially use the programmability of the random oracle to implement a non-committing encryption scheme (see [11]), by adjusting the outcome of H_{VE} to ensure that the string Z_i decrypts to the correct K_P value.

We also assume that honest users securely erase their tickets after an unlock attempt with that ticket has been made (as well as any other part of their state which no longer needs to be stored).

Theorem 5.1. Under the assumption that Π_{ENC} is a CCA2-secure encryption scheme (see Definition A.5), Π_{SIG} is a 2-move blind signature scheme (see Definition A.7) and the assumptions listed in Section 5, the protocols for SETUP, REQUEST and UNLOCK given in Sections 4.1, 4.2 and 4.3, UC-realize the ideal functionality provided in Algorithms 3 and 4 under adaptive corruptions, with erasures.

To prove the theorem, we provide a simulator Sim and prove that the resulting Ideal and Real distributions are computationally indistinguishable. Throughout, we assume that the same ticket (resp. alias) is never issued twice during a REQUEST (resp. UNLOCK) procedure in an Ideal execution with a single SID. Since each of these events occurs with at most $\lambda'^2/2^n$ probability, where λ' is the total number of tickets issued, this assumption can only reduce the adversarial distinguishing probability by at most $2 \cdot \lambda'^2/2^n$, which is negligible.

C.1 Description of Simulator Sim

Simulator Sim under adaptive corruptions of parties

Note that since we assume secure channels, Sim only needs to begin simulating the view at the moment that some party is corrupted.

Fix an environment Env, Server Server, users $\mathcal{U}_1, \ldots, \mathcal{U}_m$ and adversary $\mathcal{A}$. Recall that we allow the environment Env to choose the inputs of all parties. Simulator Sim does the following:

1. Initialization: Initialize tables $\mathcal{B}, \mathcal{E}, \mathcal{S}, \mathcal{Z}, \mathcal{T}_{\mathsf{gen}}, \mathcal{T}_{\mathsf{used}}$ to empty and counters count_i for $i \in [m]$ to 0.

2. Preprocessing: Let λ'_i be the maximum number of tickets for each party $\mathcal{U}_i$. For $i \in [m]$, $j \in [\lambda'_i]$: Generate $B^i_j \leftarrow \{0,1\}^n$, $S^i_j \leftarrow \{0,1\}^n$, $Z^i_j \leftarrow \{0,1\}^{2n}$. Add all generated B^i_j (resp. S^i_j, Z^i_j) values to $\mathcal{B}$ (resp. $\mathcal{S}, \mathcal{Z}$). Let λ' be the total number of (B^i_j, S^i_j, Z^i_j) tuples generated.

3. Responding to corruption requests:

   Corruption of a party $\mathcal{U}_i$: Sim corrupts the corresponding ideal party and obtains its internal state, consisting of unused tickets $t^i_1, \ldots, t^i_{\lambda_i}$. For $j \in [\mathsf{count}_i]$, modify entry $(U^i, S^i_j, B^i_j, E^i_j, F^i_j, Z^i_j, \perp) \in \mathcal{T}$ to $(U^i, S^i_j, B^i_j, E^i_j, F^i_j, Z^i_j, t^i_j)$. For $j \in \{\mathsf{count}_i + 1, \ldots, \lambda_i\}$:

   (a) Generate $E^i_j = \mathtt{ENC}_{PK_S}(B^i_j)$ and F^i_j as a blind signature of E^i_j using SK_S (note that since $\lambda_i - \mathsf{count}_i > 0$, Sim must have already generated $(PK_S, SK_S, PK'_S, SK'_S)$).

   (b) Add $(U^i, S^i_j, E^i_j, F^i_j, Z^i_j, t^i_j)$ to $\mathcal{T}$ and E^i_j to set $\mathcal{E}$.

   Sim releases tickets $(S^i_j, E^i_j, F^i_j, Z^i_j)$.

   Corruption of Server: Sim corrupts the corresponding ideal party and obtains its ideal internal state If an Initialize query has not yet been submitted to the ideal functionality, Sim returns $\perp$. Otherwise, if the server's keys have not yet been sampled, Sim samples $(PK_S, SK_S, PK'_S, SK'_S)$. Let $\alpha_1, \ldots, \alpha_\lambda$ be the aliases in the ideal internal state (if any). Associate each row in $\mathcal{T}_{\mathsf{used}}$ with a random alias so each entry in $\mathcal{T}_{\mathsf{used}}$ contains a value from $\{\alpha_1, \ldots, \alpha_\lambda\}$ in its final column. For $i \in [\lambda - |\mathcal{T}_{\mathsf{used}}|]$, Generate $B_i \leftarrow \{0,1\}^n$, $E_i = \mathtt{ENC}_{PK_S}(B_i)$ and F_i as a blind signature of E_i. Add all tuples $(B_i, E_i, F_i, \alpha_i)$ to $\mathcal{T}_{\mathsf{used}}$. For each row of $\mathcal{T}_{\mathsf{used}}$, release (E_i, F_i).

4. Responding to random oracle queries to $H_{\mathsf{pw}}, H_{\mathsf{KD}}$: Sim forwards the query to the oracle and forwards the response back.

5. Responding to random oracle queries to H_{VE}: Sim maintains a table $\mathcal{T}_{H_{\mathsf{VE}}}$. The table is initialized as empty. Each time $\mathcal{A}$ queries H_{VE} on input x, Sim checks the table to see if an entry of the form (x, y) appears in the table for some y. If yes, Sim returns y. Otherwise, Sim chooses a random y, adds entry (x, y) to $\mathcal{T}_{H_{\mathsf{VE}}}$ and returns y to A.

6. When responding to oracle queries, Sim also does the following:

   - **Bad Event 1:** If Server is corrupted and $\mathcal{A}$ makes a query to H_{pw} with input of the form

Camenisch et al. [7] showed that some natural non-committing encryption schemes in the programmable random oracle model can be proven secure in the UC setting, since the simulator only needs to program the random oracle at random inputs, which have negligible chance of being already queried or programmed. We anticipate that a similar argument would work for our scheme, since D^i_j is unpredictable and with very high probability will not be queried in any other session before being programmed in the target session. However, our formal proof is only for the case where an independent random oracle is assumed for each session.

$S^i_j||\hat{P}^i_j$, where $S^i_j \in \mathcal{S}$ and $(S^i_j,\cdot,\cdot,\cdot,t^i_j) \notin \mathcal{T}$ (for $t^i_j \neq \perp$) then Sim aborts.

- **Bad Event 2:** If Server is not corrupted and $\mathcal{A}$ makes a query to H_{KD} with input of the form $(B^i_j||\hat{C}^i_j)$, where $B^i_j \in \mathcal{B}$, then Sim aborts.
- If Server is corrupted and $\mathcal{A}$ makes a query to H_{pw} with input of the form $S^i_j||\hat{P}^i_j$ where $S^i_j \in \mathcal{S}$, Sim finds the tuple of the form $(S^i_j,\cdot,\cdot,\cdot,t^i_j) \in \mathcal{T}$ and submits $\textsc{Unlock}(\textsc{sid},t^i_j,\hat{P}^i_j)$ to the ideal functionality. Sim receives $(\textsc{Unlock},\textsc{sid},\alpha)$ from the ideal functionality, and returns $(\textsc{sid},\textsc{Unlock},1)$. If the ideal functionality returns $\perp$, Sim forwards $\hat{C}^i_j = H_{\mathsf{pw}}(S^i_j||\hat{P}^i_j)$ to $\mathcal{A}$. If the ideal functionality returns K_P, Sim computes $\hat{C}^i_j = H_{\mathsf{pw}}(S^i_j||\hat{P}^i_j)$, $D^i_j = H_{\mathsf{KD}}(B^i_j||\hat{C}^i_j)$ and entries for $(0||D^i_j,y_1),(1||D^i_j,y_2)$ such that $y_1||y_2 = Z^i_j \oplus (0^n,K_P))$ to $\mathcal{T}_{H_{\mathsf{VE}}}$. Sim returns $\hat{C}^i_j$ to A. **Bad Event 3:** If at this point $0||D^i_j$ or $1||D^i_j$ have already been queried to H_{VE}, Sim aborts.

7. Responding to messages from the $\textsc{Request}$ protocol issued by a corrupted $\mathcal{U}_i$ when Server is not corrupted. Sim does the following:

   (a) Generate (PK_S,SK_S,PK'_S,SK'_S) if not yet generated.

   (b) Submit $\textsc{Request}(\mathcal{U}_i,\textsc{sid},0,0)$ to the ideal functionality and receive back ticket t.

   (c) Place $(U_i,*,*,*,*,t) \in \mathcal{T}_{\mathsf{gen}}$.

   (d) Play the part of an honest signer with secret key SK'_S in the blind signature protocol with the corrupted user.

8. Responding to $(\textsc{sid},\textsc{Request},U_i)$ messages from Ideal Functionality. Sim does the following:

   (a) Set $\mathsf{count}_i := \mathsf{count}_i + 1$ and $j := \mathsf{count}_i$.

   (b) Generate $E^i_j := \mathtt{ENC}_{PK_S}(B^i_j)$.

   (c) Participate in a blind signature protocol on message E^i_j with the corrupted Server to obtain signature F^i_j.

   (d) Store $(U_i,S^i_j,B^i_j,E^i_j,F^i_j,Z^i_j,\perp) \in \mathcal{T}_{\mathsf{gen}}$.

9. Responding to messages from the $\textsc{Unlock}$ protocol issued by adversary $\mathcal{A}$ when Server is not corrupted. $\mathcal{A}$ sends $(\hat{E},\hat{F},\hat{C})$ to the server.

   - If a tuple of the form $(\cdot,\hat{E},\cdot,\hat{t},*) \in \mathcal{T}_{\mathsf{used}}$, then send $\textsc{Unlock}(\textsc{sid},\hat{t},\perp)$ to the ideal functionality.
   - Otherwise, if the signature does not verify submit $\textsc{Unlock}(\textsc{sid},\perp,\perp)$ to the ideal functionality.
   - Otherwise, if $\hat{E} = E^i_j \in \mathcal{E}$:

     (a) Find an entry of the form $(\cdot,\cdot,\hat{E},\cdot,\hat{t}) \in \mathcal{T}$. Add $(\hat{B},\hat{E},\hat{F},t,*)$ to $\mathcal{T}_{\mathsf{used}}$.

     (b) **Bad Event 4:** If there is more than one oracle query that returned $\hat{C}$, Sim aborts.

     (c) If the unique query exists, extract the password guess $\hat{P}$ (with bit length at most n'). If it does not exist, set $\hat{P}$ to $\perp$. Send $\textsc{Unlock}(\textsc{sid},\hat{t},\hat{P})$ to the ideal functionality. **Bad Event 5:** If $\hat{C} = H_{\mathsf{pw}}(S^i_j,*)$, for some $S^i_j \in \mathcal{S}$, but $\mathcal{A}$ did not make an oracle query returning $\hat{C}$, Sim aborts.

     (d) If the ideal functionality returns a value K_P, then set $D^i_j = H_{\mathsf{KD}}(B^i_j||\hat{C})$. Add $(0||D^i_j,y_1),(1||D^i_j,y_2)$ to $\mathcal{T}_{H_{\mathsf{VE}}}$ such that $y_1||y_2 = Z^i_j \oplus (0^n,K_P))$ Return D_j to $\mathcal{A}$. **Bad Event 6:** If $\mathcal{A}$ has already queried H_{VE} on $0||D^i_j$ or $1||D^i_j$, Sim aborts.

     (e) Otherwise, return $D^i_j = H_{\mathsf{KD}}(B^i_j||\hat{C}^i_j)$.

   - Otherwise if $\hat{E} \notin \mathcal{E}$, Sim does the following:

     (a) **Bad Event 7:** If there is no entry of the form $(*,*,*,*,\hat{t}) \in \mathcal{T}$, Sim aborts.

     (b) Find an entry of the form $(*,*,*,*,\hat{t}) \in \mathcal{T}$ and remove it.

     (c) Decrypt $\hat{E}$ using SK_S to obtain $\hat{B}$. **Bad Event 8:** If $\hat{B} \in \mathcal{B}$, Sim aborts.

     (d) Make an $\textsc{Unlock}$ request to the ideal functionality $\textsc{Unlock}(\textsc{sid},\hat{t},\perp)$

     (e) Continue the execution honestly to recover $\hat{D} = H_{\mathsf{KD}}(\hat{B}||\hat{C})$. Return $\hat{D}$ to $\mathcal{A}$.

10. Responding to $(\textsc{Unlock},\textsc{sid},\alpha)$ messages from Ideal Functionality. If Sim receives a message $(\textsc{sid},\textsc{Unlock},\alpha)$ (which does not stem from an $\textsc{Unlock}$ request submitted by Sim) then Sim does the following:

    (a) If there is some $(\hat{B},\hat{E},\hat{F},*,\alpha) \in \mathcal{T}_{\mathsf{used}}$. Then Sim forwards $(\hat{E},\hat{F})$ to Server, along with a random value for $\hat{C}$.

    (b) If not, update the next tuple of the form $(\hat{B},\hat{E},\hat{F},*,\perp) \in \mathcal{T}_{\mathsf{used}}$, to $(\hat{B},\hat{E},\hat{F},*,\alpha)$. Forward $(\hat{E},\hat{F})$ to Server, along with a random value for $\hat{C}$.

(c) If Sim returns $\bot$, then return 0 to the ideal functionality.

(d) Otherwise, Sim receives back a D value from Server and checks whether D was computed correctly with respect to $\hat{B}$ and $\hat{C}$. If yes, Sim sends $(\textsc{sid}, \textsc{Unlock}, 1)$ to the ideal functionality. Otherwise, Sim sends $(\textsc{sid}, \textsc{Unlock}, 0)$ to the ideal functionality. If tuples of the form $(0||D, y_1), (1||D, y_2)$ are not in $\mathcal{T}_{H_{\mathsf{VE}}}$, Sim chooses random y_1, y_2 and adds $(0||D, y_1), (1||D, y_2)$ to $\mathcal{T}_{H_{\mathsf{VE}}}$. **Bad Event 9:** If $y_1||y_2 \oplus Z = 0^n||*$, for some $Z \in \mathcal{Z}$, Sim aborts.

We begin by bounding the probability that the Bad Events occur. It is clear by inspection that Bad Event 1 occurs with probability at most $q \cdot \lambda'/2^n$, and that Bad Event 4 occurs with probability at most $q^2/2^n$, where q is the total number of oracle queries made by the adversary and Sim. Moreover, it is clear that if Bad Event 2 does not occur, then Bad Events 3 and 6 occur with probability at most $q^2/2^n$ each. We proceed to bound the remaining events (Events 2, 5, 7, 8).

Lemma C.1. Bad Event 5 occurs with at most negligible probability in the Ideal experiment.

We upper bound the probability of Bad Event 5 by analyzing the probability that $\hat{C} = H(S^i_j, x)$, for some value of $x \in \{0,1\}^{n'}$. This probability can be upper bounded by $\frac{2^{n'}}{2^n}$, since there are $2^{n'}$ possible strings of the form $S^i_j||x$ and each of these gets mapped to a particular string $\hat{C}$ with probability $\frac{1}{2^n}$. Setting parameters appropriately, we have that $\frac{2^{n'}}{2^n}$ is negligible.

Lemma C.2. Assuming the CCA2 security of encryption scheme $\texttt{ENC}$ (see Definition A.5), the probability that Bad Event 2 or Bad Event 8 occurs is at most negligible in the Ideal experiment.

The proof proceeds by showing that if Bad Event 2 or Bad Event 8 occurs with non-negligible probability, then there must be some $i \in [m]$, $j \in [\lambda'_i]$ and efficient Env, $\mathcal{A}$ (who did not corrupt Server) such that $\mathcal{A}$ queries H_{KD} on the value, B^i_j, or, in an $\textsc{Unlock}$ request, sends an encryption $\hat{E} \notin \mathcal{E}$ that decrypts to B^i_j, with non-negligible probability. We will use Env, $\mathcal{A}$ to obtain another efficient adversary $\mathcal{A}'$ who breaks the security of the CCA2 encryption scheme $\texttt{ENC}$.

The adversary $\mathcal{A}'$ breaking the CCA2 security of the encryption scheme $\texttt{ENC}$ proceeds as follows: $\mathcal{A}'$ plays the part of Sim in the Ideal experiment, with the exception that (1) It knows all the honest users passwords and keys (since it controls Env); (2) It receives PK_S externally from its CCA2 challenger (and does not know the corresponding SK_S), (3) It aborts and outputs $0, 1$ with probability $1/2$ if $\mathcal{A}$ requests a Server corruption. Sim chooses random strings $B^i_j, B'^i_j \mathcal{B}$. Upon corruption of party $\mathcal{U}_i$, $\mathcal{A}'$ Sim sends B^i_j, B'^i_j back to its CCA2 challenger. The CCA2 challenger chooses $\tilde{b} \leftarrow \{0,1\}$ and returns an encryption of B^i_j if $\tilde{b} = 0$ and an encryption of B'^i_j if $\tilde{b} = 1$. Let E^* denote the challenge ciphertext that $\mathcal{A}'$ receives in return. $\mathcal{A}'$ continues to play the part of Sim, but includes challenge ciphertext E^* in the information returned for the corruption request for party $\mathcal{U}_i$, instead of a newly generated ciphertext. When responding to $\textsc{Unlock}$ queries $(\hat{E}, \hat{F})$, Sim must decrypt using SK_S if $\hat{E} \notin \mathcal{E}$. But in this case, either (1) $\mathcal{A}'$ has not yet requested/received its challenge ciphertext from the CCA2 challenger or (2) $\hat{E} \neq E^*$, since $E^* \in \mathcal{E}$. So $\mathcal{A}'$ forwards the decryption query $\hat{E}$ to its CCA2 oracle. Recall that throughout the experiment, $\mathcal{A}'$ (playing the part of Sim) monitors all queries made to the random oracles. If an $\textsc{Unlock}$ request is made with a valid ticket that includes E^* and a $\hat{C}^i_j$ value corresponding to the correct password, $\mathcal{A}'$ chooses a value for D^i_j at random (without querying oracle H_{KD}). If, at any point, **Case 1:** a query to H_{KD} of the form $(B^i_j, *)$ is made or some CCA2 decryption oracle query yields value B^i_j, then $\mathcal{A}'$ aborts the experiment and returns 0 to its challenger. If, at any point, **Case 2:** a query to H_{KD} of the form $(B'^i_j, *)$ is made or some CCA2 decryption oracle query yields value B'^i_j, then $\mathcal{A}'$ aborts the experiment and returns 1 to its challenger. If the experiment completes without the above cases occurring, $\mathcal{A}'$ flips a coin and returns the outcome to its challenger.

Now, note that if Bad Event 2 or 8 occur with non-negligible probability $\rho = \rho(n)$, then we must have that $\Pr[\tilde{b} = 0 \wedge$ **Case 1** occurs $] = \Pr[\tilde{b} = 1 \wedge$ **Case 2** occurs $] = \rho/2$.

On the other hand, it is always the case that $\Pr[\tilde{b} = 0 \wedge$ **Case 2** occurs $] = \Pr[\tilde{b} = 1 \wedge$ **Case 1** occurs $] = q/2^{n+1} + \lambda'/2^{n+1}$, where q is the total number of distinct oracle queries made during the experiment. This is because when $\tilde{b} = 0$, there is no information at all about B'^i_j contained in adversary $\mathcal{A}$'s view (unless $B'^i_j = B^{i'}_{j'}$ for some $(i', j') \neq (i, j)$, which occurs with probability at most $\lambda'/2^{n+1}$) and so $\mathcal{A}$ can only happen to query the oracle on B'^i_j at random. The case for $\tilde{b} = 1$ follows by identical reasoning.

Thus, the distinguishing advantage of CCA2 adversary $\mathcal{A}'$ is $\rho/2 - q/2^{n+1} - \lambda'/2^{n+1}$, which is non-negligible, since ρ is non-negligible. This implies a contradiction to the CCA2 security of the underlying encryption scheme.

Lemma C.3. Assuming the unforgeability of the blind signature scheme (see Definition A.7), Bad Event 7 occurs with at most negligible probability in the Ideal experiment.

The proof proceeds by showing that if Bad Event 7 occurs with non-negligible probability for some efficient adversary A, then, by definition, we obtain an efficient adversary $\mathcal{A}'$ who submits a larger number of valid UNLOCK requests than there are valid tickets obtained from the ideal functionality. But note that each valid UNLOCK request is accompanied by a fresh blind signature $\hat{F}$. Moreover, the number of valid signatures obtained from the signer corresponds to the number of valid tickets obtained. Thus, adversary $\mathcal{A}$ can be used to obtain adversary $\mathcal{A}'$ such that, according to Definition A.7, breaks the security of the blind signature scheme.

Conditioned on the Bad Events not occurring, the only difference between a Real and Ideal execution, is that in the Ideal execution in Step (10b) the simulator submits the next available $(\hat{E}, \hat{F})$ pair, whereas in the Real execution the order of submitted $(\hat{E}, \hat{F})$ pairs depends on which party is making the UNLOCK request. However, the blindness property of the blind signature scheme ensures that given a set of interactions and message signature pairs, the signer cannot tell in which order the message signature pairs were generated. Indeed, this is the case even when (PK_S', SK_S') are adversarially generated. Thus, the view of the adversary is indistinguishable in the two cases. We therefore conclude with the following lemma.

Lemma C.4. Assuming the blindness of the blind signature scheme (see Definition A.7), the Ideal and Real output distributions are computationally indistinguishable.

Optimization of a magnetostatic cavity for a ^{3}He spin analyzer on the CANDOR polychromatic reflectometer

Md. T. Hassan[1,2], **W.C. Chen**[1,2,*]

[1] National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA
[2] University of Maryland, College Park, Maryland 20742, USA

E-mail: wcchen@nist.gov

Abstract.

A new instrument, the CANDOR (Chromatic Analysis Neutron Diffractometer or Reflectometer) polychromatic beam reflectometer is being built at the National Institute of Standards and Technology Center for Neutron Research. CANDOR will have a capability of polarization analysis for both specular and off-specular scattering, and allow for measuring the magnetic structures of magnetic thin films with a significantly shortened measurement time as compared to a conventional monochromatic beam reflectometer at a continuous neutron source. A critical component of the polarization analysis capability is to develop a ^{3}He neutron spin filter (NSF) for polarization analysis for both specular and off-specular scattering for a continuous wavelength range between 4 Å and 6 Å and a large solid angle coverage for up to 20 energy dispersive detector channels. For *ex-situ* spin-exchange optical pumping (SEOP) operation, a magnetically shielded solenoid (MSS) will be employed to provide a volume averaged field gradient, $|\vec{\nabla} B_\perp/B|$, better than 5×10^{-4} cm^{-1} to maintain ^{3}He polarization relaxation times of at least a few hundred hours on the beam line even when there are significant stray fields from a magnet operating at a field up to 3 T at the sample position. We describe a procedure to design, construct, and optimize a MSS with non-identical compensation coil dimensions on the ends to match the scattered neutron beam rays. The volume averaged field gradient is quantified and optimized over a cell for a MSS using the finite-element software package RADIA and compared with that from the experiment.

Key words: polychromatic reflectometer, CANDOR, SEOP, ^{3}He neutron spin filter, field gradient

1. Introduction

A conventional monochromatic reflectometer at a continuous source utilizes either a pyrolytic graphite (PG) crystal or a velocity selector to monochromatize the incident beam and detects neutrons elastically scattered from the sample without energy analysis. At the National Institute of Standards and Technology (NIST) Center for Neutron Research a novel instrument, the CANDOR (Chromatic Analysis Neutron Diffractometer or Reflectometer) polychromatic beam reflectometer is being built and is expected to be commissioned soon. CANDOR utilizes a polychromatic beam with incident neutron wavelengths (energies) between 4 Å (5.113 meV) and 6 Å (2.272 meV) that are closely matched to the peak flux of the neutron spectra from the NCNR's cold neutron source [1]. The energy determination is accomplished at the scattered

Hassan, Md.; Chen, Wangchun. "Optimization of a Magnetostatic Cavity for a u3He Spin Analyzer on the CANDOR Polychromatic Reflectometer." Paper presented at 12th International Conference on Polarised Neutrons for Condensed Matter Investigations, Milton Hill, GB. July 03, 2018 - July 06, 2018.

beam side using a linear array of PG crystals, similar to the Continuous Angle Multiple Energy Analysis (CAMEA) concept of the multi-analyzer module MultiFlexx of the cold triple axis spectrometer FLEXX [2]. The analyzer system employs a bank of 18 energy dispersive detector channels, each covering an angular range of 0.355 °, each containing 54 graphite analyzer crystals in series with corresponding scintillation detector plates. The total angular coverage for the scattered beam will be 6.4 °. The neutron detection system after the PG crystals is one of the most critical units for the CANDOR development and has been accomplished using a slab of LiF:ZnS(Ag) scintillator with embedded wavelength shifting fibers and a silicon photomultiplier used as a photosensor [3]. It is expected that CANDOR will provide significantly higher usable intensity incident on sample, hence reducing measurement times significantly for a given sample size over a conventional continuous source-based monochromatic reflectometer.

CANDOR will be equipped with a longitudinal polarization analysis capability for both specular and off-specular reflectivity measurements. For the polarized beam configuration, a double-V supermirror made by SwissNeutronics [4] is used to polarize the incident neutron beam and is expected to provide a neutron polarization of 98 % or higher for the entire wavelength range from 4 Å to 6 Å. An adiabatic radio-frequency (RF) spin flipper located immediately following the monochromator is used to flip the neutron spin over a large cross sectional beam with a flipping efficiency better than 99.5 % [5]. To spin analyze the neutron polarization of the scattered beam, it is necessary to have a spin analyzer able to cover the entire 6.4 ° angular range for the entire energy range and provide a spatially uniform analyzing power higher than 98 %, and have a second broadband spin flipper with a flipping efficiency higher than 99.5 % over a large cross sectional scattered neutron beam. Given the space of 60 cm available for both the spin analyzer and the broadband spin flipper, a ^{3}He neutron spin filter (NSF) was chosen for the following reasons. A ^{3}He NSF can (1) polarize a broad wavelength band of neutrons effectively; (2) polarize large area and widely divergent neutron beams; and (3) efficiently flip the neutron polarization by reversing the ^{3}He nuclear polarization using the adiabatic fast passage (AFP) nuclear magnetic resonance (NMR) technique [6], hence integrating the analyzer and flipper into a single device. The advantage of integration of the analyzer and the broadband flipper into a single neutron polarizing device is a unique feature of the ^{3}He analyzer since no other conventional neutron spin analyzer and flipper together can be fit into a 60-cm space along the beam path for the CANDOR scattered beam configuration. At the NCNR, ^{3}He NSFs have been routinely applied to small-angle neutron scattering, triple-axis spectrometry, and wide-angle polarization analysis [7]. A ^{3}He NSF has been demonstrated previously for polarization analysis of the diffusely reflected neutrons on the polarized neutron reflectometer at the NCNR [8, 9].

For ^{3}He NSF applications in polarized neutron scattering, a cylindrical geometric shape is generally chosen for a ^{3}He cell. For the CANDOR analyzer, a cylindrical ^{3}He cell with a diameter of 13 cm or larger is necessary for covering the scattering angle subtended by the entire detector bank. For this large ^{3}He cell, we plan to operate the ^{3}He NSF in the *ex-situ* spin exchange optical pumping (SEOP) operation mode. The ^{3}He gas is polarized in sealed cells in an off-line lab using the SEOP method, and the cells are transported to the beam line in a portable solenoid. In the *ex-situ* SEOP operation mode, it is critical to maintain the relaxation time of nuclear-polarized ^{3}He gas as long as possible. The relaxation time T_1 of the polarized ^{3}He gas has three contributions 1) dipole-dipole interactions [10], T_1^{dd}, 2) wall relaxation [11], T_1^{w}, 3) magnetic field gradients [12], T_1^{fg}, and is given by

$$\frac{1}{T_1} = \frac{1}{T_1^{\mathrm{dd}}} + \frac{1}{T_1^{\mathrm{w}}} + \frac{1}{T_1^{\mathrm{fg}}} \tag{1}$$

The dipole-dipole relaxation rate increases linearly with the partial pressure of ^{3}He gas and has been determined to be $p/807$ (h^{-1}) for p in bar [10]. The relaxation rate due to magnetic-field gradients is directly proportional to the square of the fractional transverse field gradients and inversely proportional to the ^{3}He pressure [12]. As discussed in Eq. 2 later, for a cell at a pressure of one bar in a uniform gradient of $|\vec{\nabla}B_\perp/B|$=5 × 10^{-4} cm^{-1} the gradient-induced relaxation time is ≈600 h. For a given ^{3}He polarization of 85 %, which can be routinely achieved at NIST with the SEOP method using the spectrally narrowed high power diode bar lasers [13], it is desired to have a ^{3}He partial pressure of ≈1 bar for the CANDOR ^{3}He analyzer for optimizing the analyzing power and transmission. This implies a dipole-dipole relaxation time of ≈800 h. The current best ^{3}He cell fabrication practice has nearly eliminated the wall relaxation rate (with the wall relaxation time longer than 3000 h) in certain ^{3}He cells [14, 15]. Specifically we have successfully developed ^{3}He cells for CANDOR with relaxation times between 450 h and 690 h, measured under a homogeneous magnetic field with the field gradient $|\vec{\nabla}B_\perp/B|$ better than 1×10^{-4} cm^{-1}. For the best cell, this implies a wall relaxation time longer than 5000 h. In order to achieve a relaxation time of 250 h or longer for the ^{3}He analyzer, a volume averaged field gradient upper limit of 5×10^{-4} cm^{-1} is required.

The polarization analysis capability on CANDOR will likely use a hybrid 3 T superconducting magnet at the sample with the field along the vertical direction. In order to effectively shield stray fields from the 3 T magnet, we will employ a magnetically shielded solenoid (MSS) to maintain the ^{3}He polarization on the CANDOR beam line. Several different magnetically shielded solenoids have been developed to maintain the ^{3}He polarization for several polarized beam instruments at the NCNR, and are able to shield stray fields of approximately a few mT, produced from either an electromagnet or even a superconducting magnet [7]. Furthermore, it is necessary to implement a neutron spin rotation device to adiabatically rotate the spin from the vertical to longitudinal direction. Under such an operating configuration, it is still necessary to maintain volume average field gradient better than 5×10^{-4} cm^{-1}. In this paper we describe modeling of the field gradients, design and construction of the CANDOR MSS, and characterization of field gradients for the ^{3}He spin analyzer on CANDOR.

2. Modeling

Similar to our previous designs of solenoids [7, 9, 16, 17, 18], the 16 AWG (American Wire Gauge) 1.31 mm diameter copper wire was wound on an aluminum cylinder which was nested inside a Conetic mu-metal [19] cylinder. The gap between the two cylinders is typically 7 mm. On each end of the mu-metal solenoid, a Conetic mu-metal end cap is tightly attached to the mu-metal cylinder body with an overlap of 2.5 cm. The overlap protects the mu-metal from saturation even when applying a modest current to the aluminum solenoid. A hole in each end cap lets neutrons pass through without neutron depolarization. Fig. 1 shows a schematic of the CANDOR MSS. Notable features in Fig. 1 include a RF coil structure that is used to invert the ^{3}He polarization, compensation coils centered to each end cap hole, and borated aluminum neutron shields attached to both the upstream end cap and the downstream compensation coil frame. We began by determining the dimensions of the CANDOR MSS with the aim of designing the solenoid to be as compact as possible, while still achieving a volume averaged field gradient better than 5×10^{-4} cm^{-1}. Based on the ^{3}He cell diameter requirement, we determined the diameters of the aluminum cylinder and mu-metal cylinder to be 261 mm and 279 mm, respectively. The length of the mu-metal solenoid was determined to be 356 mm by taking the following facts into consideration: (1) the total space available, (2) the space necessary for a device to accomplish an adiabatic $\pi/2$ spin rotation, (3) the effect of the magnetic field from the adiabatic spin rotation device on the field gradient of the MSS. Unlike magnetically shielded solenoids developed earlier at the NCNR [7] where the same number of ampere-turns in two identical compensation coils have been used, we used compensation coils with different diameters and hence different number of ampere-turns in order to further improve the field gradient and/or make it more compact for the CANDOR MSS. The diameters for the compensation coils were determined to be 117 mm and 143 mm attached to the upstream and downstream end caps, respectively. The compensation coils were wound on a 5.8 mm wide and 4 mm deep aluminum groove structure that was attached to the mu-metal end cap as shown in Fig. 1. The groove structure displaced the compensation coils by 3 mm along z and 2.8 mm in the radial direction from the end cap hole. As discussed later, the field gradient induced relaxation rate with non-identical compensation coils will be reduced by a factor of 1.88 over the conventional identical compensation coils. An additional benefit of using non-identical compensation coils is a possibility of reducing the diameter of the ^{3}He analyzer cell since the field homogeneity center is shifted closer to the upstream end cap as described later.

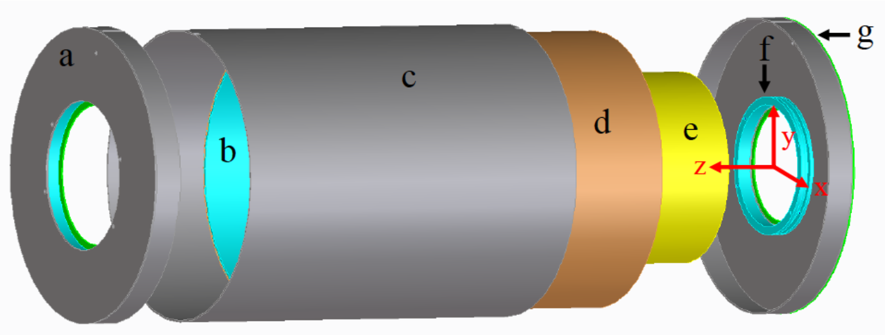


Figure 1. (Color online) Schematic of a magnetically shielded solenoid with circular holes in the end caps. The MSS is 279 mm in diameter, 356 mm long, with different sizes (117 mm diameter and 143 mm diameter) of holes on the end caps as described in the text. The MSS has the following components, (a) Conetic [19] mu-metal end caps, (b) aluminum cylinder with a wall thickness of 2.4 mm for support of the copper winding, (c) Conetic [19] mu-metal cylinder, (d) copper winding, (e) RF coil structure, (f) compensation coils centered to the end cap hole, (g) borated aluminum neutron shielding piece attached to the end cap in the upstream direction. The thickness for all mu-metal pieces is 1.6 mm.

We proceed to discuss the simulation of the magnetic field profile, the resulting field gradients, and optimization of compensation coils of the CANDOR MSS. A similar simulation and modeling has been completely described for our earlier development of a compact end-compensated magic box [20]. The field gradients relevant to the relaxation of polarized ^{3}He gas are $\vec{\nabla} B_x$ and $\vec{\nabla} B_y$, using the coordinate system in Fig. 1. These six gradients of field components, all perpendicular to the applied field, contribute over the volume, V, of a cell to the relaxation time of the polarized ^{3}He gas. At room temperature, the relaxation rate $1/T_1^{\text{fg}}$ due to field gradients is given by

$$\frac{1}{T_1^{\text{fg}}} = \frac{6700}{pV} \iiint_V \left(\frac{|\vec{\nabla} B_x|^2}{B^2} + \frac{|\vec{\nabla} B_y|^2}{B^2} \right) dx\,dy\,dz \;\; \text{h}^{-1} \equiv \frac{6700}{p} |\vec{\nabla} B_\perp / B|^2 \;\; \text{h}^{-1} \qquad (2)$$

where p is the gas pressure in bar and $\frac{\vec{\nabla} B_x}{B}$ and $\frac{\vec{\nabla} B_y}{B}$ are the gradients in the transverse components of the magnetic field (for nuclear polarization along the z-axis) normalized to the central field, B, in units of cm^{-1}[21]. $|\vec{\nabla} B_\perp / B|$ is the normalized volume-averaged transverse gradient over the cell volume, V. In practice, these transverse components are too tiny ($< 0.1\mu$T) to measure experimentally for a field gradient level of 10^{-4} cm^{-1} using conventional means. As demonstrated in an earlier publication [20], the conveniently measurable gradient component along the applied field $|\partial B_z / \partial z| / B$ was a good indicator of the normalized volume-averaged transverse field gradient.

We used the finite-element software package RADIA [22] for the field profile and the Mathematica [23] interface for analytical calculation of the magnetic field gradients. We modeled and optimized the CANDOR MSS with three different configurations: (i) the upstream compensation coil with an inner diameter of 117 mm and downstream compensation coil with an inner diameter of 143 mm, both upstream and downstream compensation coils with a diameter of 117 mm (ii) and 143 mm (iii). The criterion for convergence of the RADIA simulation iterations was that the magnetization of all segments changed by less than 10 nT. The segment size was determined to be about 4.6 mm (78 sections) along z and 6.67 ° (54 sections) radially in the x-y plane by confirming the normalized volume-averaged field gradients changed by less than 10^{-5} cm^{-1}.

We began to model the MSS with two identical compensation coils with an inner diameter of 117 mm and 143 cm as it is more straightforward and has been done for earlier development of the MSSs [7]. For the simulations, no air gap was used between the mu-metal cylinder body and each end cap in the 2.5 cm overlap section. The main coil winding section on the aluminum cylinder was extended to be in contact with each mu-metal cap. The compensation coil winding was fixed at 3 layers and 4.5 turns per layer in the groove. Changing the compensation coil was done via a change of the current dennsity. This closely matched to the experimentally field mapped condition since varying the number of the compensation coils was impractical during mapping. Other geometric dimensions for the simulation are identical to that from Fig. 1. The field in the center of the solenoid was fixed at about 2.73 mT. We found that at the optimized conditions the number of the compensation coils were 12 turns and 16 turns, corresponding to a ratio of ampere-turn of 0.044 and 0.059 between the compensation coil of an inner diameter of 117 mm and the main coil, and between the compensation coil of an inner diameter of 143 mm and the main coil, respectively. The volume averaged gradients $|\vec{\nabla} B_{\perp}/B|$ were calculated over a cylindrical ^{3}He cell, 120 mm in diameter and 100 mm long with a cylindrical axis along z. These gradients were determined to be 2.6×10^{-4} cm^{-1} and $4.8\ \times10^{-4}$ cm^{-1}, for compensation coils of inner diameters 117 mm and 143 mm, respectively. Although the volume averaged field gradient for the MSS with the compensation coil of an inner diameter of 143 mm was already better than $5\ \times10^{-4}$ cm^{-1} and was acceptable, it is expected that the field gradient for the MSS with non-identical compensation coils as shown in Fig. 1 would be even smaller. The simulation of the MSS with identical compensation coils gave a good indicator for the simulation of the CANDOR solenoid with non-identical compensation coils as the simulation for the MSS with non-identical compensation coils requires more complicated optimization of each individual compensation coil.

To optimize the field homogeneity of the CANDOR MSS with non-identical compensation coils, we began to tune the smaller compensation coil while fixing the larger compensation coil at 16 turns as determined from simulation with identical compensation coils. The optimal condition was determined to be 16 turns for the smaller compensation coil. The smaller compensation coil was then set to 16 turns and the procedure was repeated for the larger compensation coil. The optimal condition was determined to be 19 turns for the larger compensation coil. The volume averaged field gradient was calculated in every step. The above procedure was reiterated until the calculated volume averaged field gradient changed by less than 10^{-5} cm^{-1}. Fig. 2 shows an example of the calculated field B_{z} profiles by varying the number of turns of the smaller compensation coil while fixing the larger compensation coil to 19 turns. It appeared that 16 turns of the smaller compensation coil yielded the smallest volume averaged transverse field gradient.

Fig. 3 shows contour plots of the simulated B_{z} in the central plane (x=0). Shown in Fig. 3b is a contour plot zoomed from Fig. 3a centered around the most homogenous field region. The field in the center was about 2.73 mT and chosen to not saturate the mu-metal, provide a reasonably high field so that the effect from any external field to the field inside the MSS is minimal, and to achieve a very efficient AFP inversion of the ^{3}He polarization. The effect of compensation coils

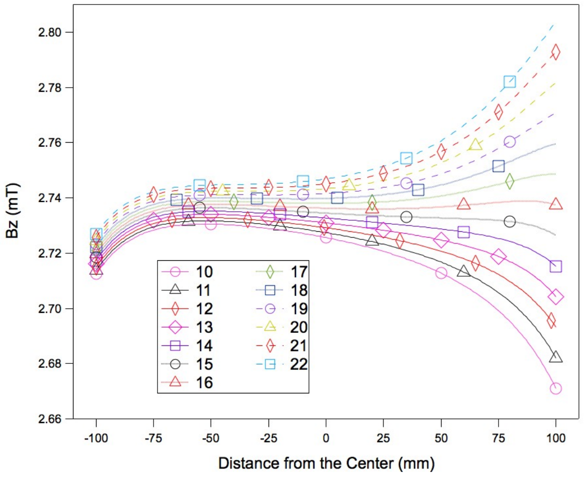


Figure 2. (Color online) Calculated B_z field profiles for different numbers of turns of compensation coils centered on the 117 mm diameter hole end cap with the compensation coil centered on the 143 mm diameter hole fixed to 19 turns. Lines are to guide the eye.

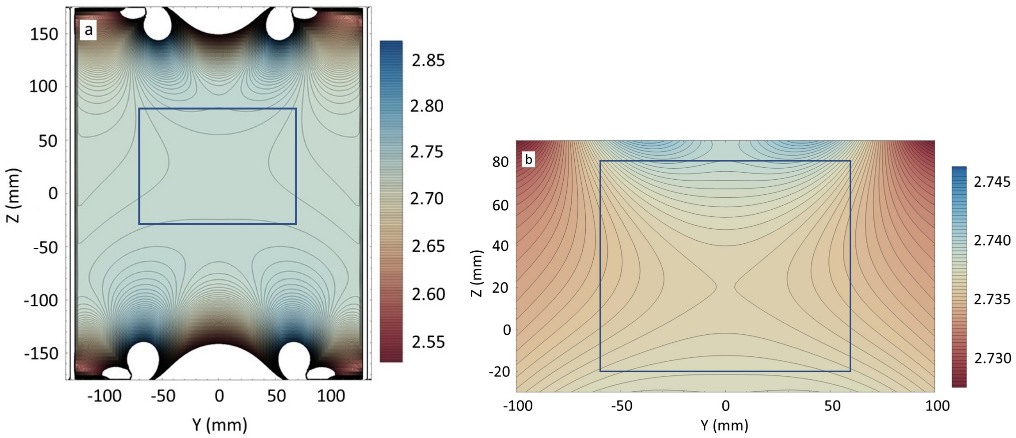


Figure 3. (Color online) Contour plots of the calculated B_z profile in the central plane (x=0) for the full size solenoid (a) and for the zoomed part centered around the most homogenous field region (b) at the optimized configuration of 16 turns and 19 turns of compensation windings attached to the small and large end cap hole, respectively. The field is in the unit of mT. The boxes in Fig. 3 indicates where the ^{3}He analyzer cell should be positioned.

Hassan, Md.; Chen, Wangchun. "Optimization of a Magnetostatic Cavity for a u3He Spin Analyzer on the CANDOR Polychromatic Reflectometer." Paper presented at 12th International Conference on Polarised Neutrons for Condensed Matter Investigations, Milton Hill, GB. July 03, 2018 - July 06, 2018.

was visible in Fig. 3a around the end cap holes and no magnetic flux saturation was observed near the main and compensation coils. The boxes in Fig. 3 represent a region with a minimal volume averaged field gradient and a region where the ^{3}He analyzer cell should be located. The field homogeneity center is shifted about 30 mm along z, as the center is expected at z=0 for a solenoid with identical compensation coils and holes in the end caps. In Fig. 3a, the two most inner magnetic flux lines represent a field difference of 2.2 μT, corresponding to a fractional field change $|\Delta B_z|/B$ of no worse than $8{\times}10^{-4}$ over a cross sectional area of 150 mm by 120 mm. In Fig. 3b the two most inner flux lines represent a field difference of 0.27 μT, corresponding to a fractional field change $|\Delta B_z|/B$ of $1{\times}10^{-4}$. The region of good field homogeneity with $|\Delta B_z|/B \approx 1{\times}10^{-4}$ was about 70 mm along the z-axis and nearly 120 mm along the y-axis. We calculated the volume averaged field gradient $|\vec{\nabla} B_\perp/B|$ over the volume of a cylindrical ^{3}He cell, 120 mm diameter and 100 mm long, and obtained $|\vec{\nabla} B_\perp/B|\approx 3.5{\times}10^{-4}$ cm^{-1} at the optimized configuration with ampere-turn ratios of 0.059 between the smaller compensation coil and the main coil and 0.070 between the larger compensation coil and the main coil, respectively. This gradient was a factor of 1.37 smaller than that for the MSS with the identical larger compensation coils of a diameter of 143 mm, implying an improvement of a factor of 1.88 in the field gradient induced relaxation time. This improvement of the field gradient was obtained without sacrificing the scattering angle coverage. At the optimized condition, simulations yielded 16 turns and 19 turns of 16 AWG wire for the smaller and larger compensation coils, respectively.

After obtaining an improved field gradient, we explored more contour maps for the off-center regions along both x and z directions. It was observed that the calculated $B_{\rm z}$ field homogeneity is still good at x=60 mm that corresponds to the edge of the cell and is only slightly worse than that in the center. Fig. 4 shows contour plots of the calculated $B_{\rm z}$ profile along the radial direction at z=30 mm and at z=80 mm. In Fig. 4 (a) and (b), the two most inner flux lines represented a field difference of 0.27 μT, corresponding to a fractional field change $|\Delta B_z|/B$ of $1{\times}10^{-4}$.

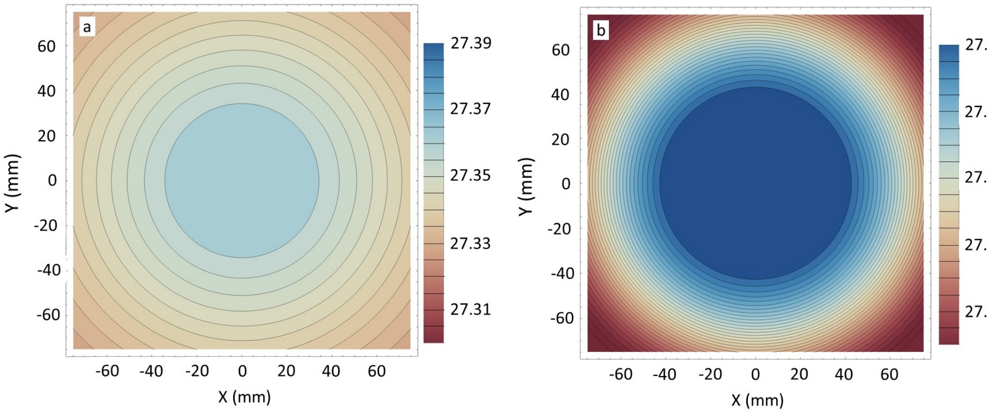


Figure 4. (Color online) Contour plots of the calculated $B_{\rm z}$ profile along the radial direction at z=30 mm (a) and at z=80 mm (b). The field is in the unit of mT.

3. Experimental results and discussion

We have designed and constructed a MSS based on modeling with the dimensions shown in Fig. 1. Each compensation coil was wound on an aluminum frame that was centered around the hole of the corresponding end cap. Both the main and compensation coils used 16 AWG wire. The main coil consisted of 270 turns. A computer controlled mapping system was used for field mapping at different compensation coil conditions. A 3-axis Hall probe (Lakeshore 460 3-channel Gaussmeter with a HSE probe) and a 3-axis fluxgate magnetometer (Bartington MG-03MS) were used to measure the field. The fluxgate has a resolution of 1 nT and was used for measurements of magnetic fields lower than 1 mT. The hall probe has a resolution of 10 nT and was used for measurements of fields higher than 1 mT. The mapper could translate along both y and z directions. Many field maps were taken on the axis of the MSS with varying turns for both compensation coils as discussed in the modeling section. Fig. 5 shows an example of a set of measured B_{z} field profiles in the central plane x=0, which was in good agreement with the simulation. We measured the line averaged gradient along z over a 100 mm length centered at the most homogenous region for each field map and found that the line-averaged gradient $|\Delta B_z/\partial z|/B$ was minimized at 20 turns of compensation coils on the 143 mm diameter hole and 11 turns of compensation coils on the 117 mm diameter hole. At this configuration, the linearly averaged gradient was 9.3×10^{-5} cm^{-1}. Off-center axis field maps were also taken to check the field homogeneity at the different compensation coil configurations. The experimentally determined compensation coil configuration at the 143 mm end cap was in good agreement with the model. However it was 11 turns of compensation coil at the 117 mm end cap, somewhat deviated from the simulation. Any tiny air gap between the Conetic mu-metal cylinder and end cap occurring during fabrication would cause perturbation of the magnetic field, while no gap was assumed for the simulation. We have checked the effect of the air gap between the Conetic mu-metal cylinder and end cap for two conditions and found an air gap of 1 mm (0.5 mm) can change the optimized configuration for the smaller compensation coil from 16 turns with no air gap to 6 (10) turns. It was determined that an $\sim$ 0.35 mm air gap would be required to change the optimized smaller compensation coil from 16 turns to 11 turns. This would result in a nearly negligible change of the volume averaged gradient $|\vec{\nabla} B_\perp/B|$ (from 3.5×10^{-4} cm^{-1} to 3.6×10^{-4} cm^{-1}). In practice, each mu-metal end cap was designed to be snugly fitted onto the mu-metal cylinder body to allow for convenient and quick ^{3}He cell exchange (on the order of several seconds) to minimize ^{3}He polarization loss. It is likely that a tiny gap exists due to the fabrication process. We speculate this air gap is likely the main reason for the discrepancy between the experiment and simulation. However, other non-ideal geometric factors such as a small difference of compensation coil location, inhomogeneous permeability in the mu-metal, roundness of the mu-metal cylinder and end caps, and parallelism of the end cap might make additional contributions to the discrepancy.

After optimization of the CANDOR MSS, we measured the relaxation time of the CANDOR ^{3}He analyzer cell Sundown. Sundown has a diameter of 132 mm and a neutron path length of 92 mm, and is able to cover the scattering angle from the entire detector assembly. Sundown was filled to a partial ^{3}He pressure of 0.95 bar. We obtained a relaxation time of 690 h in a pair of large Helmholtz coils that yielded a field gradient $|\vec{\nabla} B_\perp/B|$ better than 1×10^{-4} cm^{-1}, corresponding to a field gradient induced relaxation time of about 14000 h [20]. The relaxation time measurement was done in the CANDOR MSS after optimization on the PHADES (Polarized ^{3}He And Detector Experiment Station) beam line at the NCNR. The ^{3}He polarization was determined using the neutron transmission method [24]. Fig. 6 shows the time-dependent ^{3}He polarization. The ^{3}He polarization decays exponentially with time and is given by $P_{\mathrm{He}}(t) = P^0_{\mathrm{He}} exp(-t/T_1)$, where P^0_{He} is the initial ^{3}He polarization. We obtained a ^{3}He polarization relaxation time of 414.6 h ± 2.4 h in Fig. 6. The extrapolated volume averaged field gradient $|\vec{\nabla} B_\perp/B|$ is $(3.7\pm0.2)\times10^{-4}$ cm^{-1}, in good agreement with that from the simulation. The initial ^{3}He polarization was 85 %. The final time averaged ^{3}He polarizations would 82.6 %, 80.3 %, and 78.0 % for experiments of 24 h, 48 h and 72 h duration, respectively.

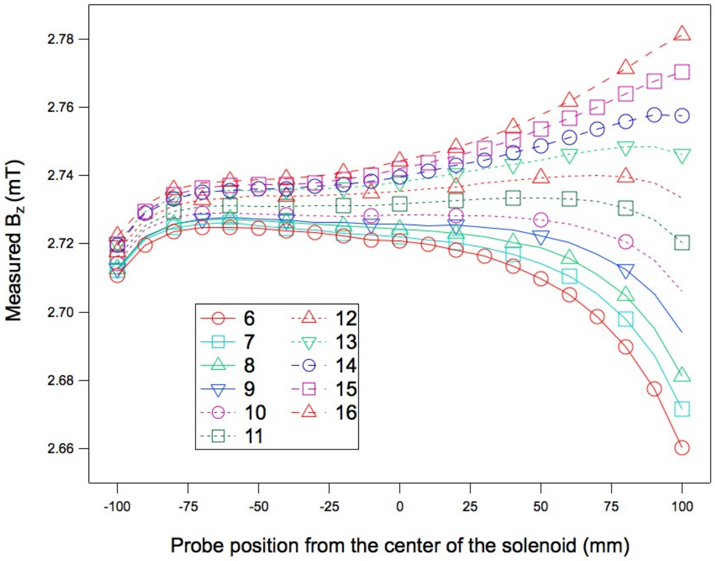


Figure 5. (Color online) Measured B_z field profiles for different numbers of turns of compensation coils centered on the 117 mm diameter hole end cap with the compensation coil centered on the 143 mm diameter hole fixed to 20 turns. The hall probe was used for these measurements. Lines are to guide the eye. Error bars are smaller than the data points.

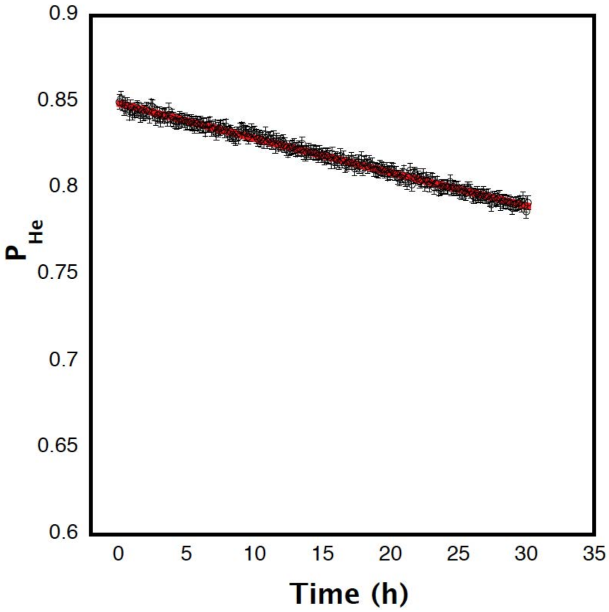


Figure 6. (Color online) The ^{3}He polarization P_{He} decay with time for the cell Sundown that will be used on CANDOR after optimization of the CANDOR MSS. The red solid line is the best fit to an exponential decay of ^{3}He polarization, yielding a relaxation time T_1 of 414.6 h ± 2.4 h. Error bars and uncertainties represent one standard deviation.

4. Conclusions

We have simulated and optimized the magnetic field gradients of a magnetically shielded solenoid (MSS) for the ^{3}He neutron spin filter that will be used as a spin analyzer on the new polychromatic reflectometer CANDOR at the NCNR. We have calculated the volume-averaged transverse field gradient which is directly linked to the field gradient-induced ^{3}He polarization relaxation time. Based on the simulation, we have constructed and characterized a compact MSS that has a dimension of 279 mm in diameter and 356 mm long. The MSS has a Conetic mu-metal cylindrical solenoid with an end cap attached to either end. By applying non-identical compensation coils centered to the hole on each end cap, we have found that the center where the field is most homogenous shifted ≈30 mm closer to the end with a smaller compensation coil. Hence the ^{3}He analyzer cell should be placed closer to the sample, which allows its diameter to be reduced. We have also shown that the volume-averaged field gradient over the cell was reduced by a factor of 1.37, hence the field gradient induced relaxation time was improved by a factor of 1.88 compared to that with identical compensation coils. We have obtained a ^{3}He polarization relaxation time of 414.6 h ± 2.4 h for a ^{3}He cell 132 mm in diameter and 92 mm long after optimization of the CANDOR MSS. After correction to the intrinsic relaxation time of 690 h, the volume averaged field gradient was determined to be $(3.7 \pm 0.2) \times 10^{-4}$ cm^{-1}, agreeing well with the calculation (3.5×10^{-4} cm^{-1}). In the near future, we will develop a ^{3}He adiabatic fast passage NMR flipper that is integrated into the ^{3}He analyzer. We plan to characterize the overall polarization analysis capability as soon as CANDOR is available.

Acknowledgments

We thank Jeff Anderson and Jack Fuller of the NIST glassblowing and optical shops. We thank Ross Erwin for his help with the RADIA simulation software and Shannon Watson for her help with construction of the magnetic shielded solenoid. Discussions with Thomas Gentile are gratefully acknowledged. The work utilized facilities supported in part by the National Science Foundation under Agreement No. DMR-1508249.

References

[1] https://ncnr.nist.gov/equipment/msnew/ncnr/candor.html
[2] Groitl, F *et al.*, 2017 *Sci. Rep* **7** 13637
[3] Osovizky, A *et al.*, 2018 *J. Phys. Commun.* **2(4)**, 045009.
[4] SwissNeutronics AG, Bruehlstrasse 28, CH-5313 Klingnau, Switzerland. Certain trade names and company products are mentioned in the text or identified in an illustration in order to adequately specify the experimental procedure and equipment used. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the stated purpose.
[5] A special RF flipper has been developed to flip the incident neutron spin with a flipping efficiency better than 0.999 for the entire 6 cm by 15 cm beam. The results will be published elsewhere.
[6] Abragam A 1961 *The Principles of Nuclear Magnetism* Oxford University Press, Oxford, England
[7] Chen WC *et al.* 2014 *Journal of Physics: Conference Series* **528** 012014
[8] Chen WC *et al* 2003 *Physica B* **335** 196
[9] Chen WC, Gentile TR, O'Donovan KV, Borchers JA, Majkrzak CF 2004 *Rev. Sci. Instrum.* **75** 3256
[10] N.R. Newbury *et al.*, 1993 *Phys. Rev. A* **48** 4411
[11] W.A. Fitzsimmons, L.L. Tankersley, and G.K. Walters, 1968 *Phys. Rev.* **179** 156
[12] L.D. Schearer and G.K. Walters, 1965 *Phys. Rev.* **139**, A1398-A1402
[13] Chen WC *et al.* 2014 *J. Appl. Phys.* **116** 014903
[14] Chen WC *et al.* 2011 *Journal of Physics: Conference Series* **294** 012003, and references therein
[15] Salhi Z, Babcock E, Pistel P, and Ioffe A 2014 *Journal of Physics: Conference Series* **528** 012015
[16] Gentile TR *et al* 2005 *Physica B* **356** 96
[17] Chen WC *et al* 2007 *Physica B* **397** 168
[18] Chen WC *et al* 2009 *Physica B* **404** 2663

[19] Magnetic Shield Corporation, 740 N. Thomas Drive, Bensenville IL 60106. Co–Netic mu-metal has a higher magnetic permeability than that of Netic mu-metal.

[20] McIver J W, Erwin R, Chen W C and Gentile T R 2009 *Rev. Sci. Instrum.* **80** 063905

[21] Cates G D, Schaefer S R and Happer W 1988 *Phys. Rev. A* **37** 2877

[22] Chubar O *et al.*, Proceedings of EPAC 2004, Lucerne, Switzerland, p. 1675.

[23] http://www.wolfram.com/mathematica/.

[24] Jones G L *et al.* 2000 *Nucl. Instrum. Meth. A* **440** 772

**Proceedings of the ASME 2019
International Design Engineering Technical Conferences
and Computers and Information in Engineering Conference
IDETC/CIE2019
August 18-21, 2019, Anaheim, CA, USA**

IDETC/CIE2019-97134

On Practice and Theory of Constructive Composite Geometry and Topology

Thomas D. Hedberg, Jr.
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, U.S.A.

Allison Barnard Feeney
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, U.S.A.

Vijay Srinivasan
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, U.S.A.

Abstract

This paper synthesizes a theory from industrial best practices codified in recent standards. Recent editions of ASME and ISO standards codify the evolving industrial best practices in defining and modeling the information about products made from fibrous composite materials. A theory of constructive composite geometry and topology is synthesized from these practices. Major features of this theory include (1) a constructive composite geometry tree that is equivalent to the ply/laminate tables of the standards, (2) an adjacency graph that captures a crucial aspect of the topology of the geometric cell complex structure of composite products, and (3) conformal mapping of ply surfaces using rosettes in the lay-up process. It also addresses the geometrical and topological structure of fiber arrangements inside the plies. The goal of the theory is to provide a scientific basis for standards that enable the digital transformation of composite product manufacturing.

1 Introduction

Products made of fibrous composite materials are everywhere, but a good theory for representing them in practice is notably absent. This has forced engineers to depend on drawings, cross-sectional views, notes, and tabular entries to define fibrous composite products and guide their manufacturing processes. Some solid and geometrical modeling tools have been improvised to aid and assist this cumbersome exercise, but these are viewed merely as workarounds in the absence of a good theory and computational tools that support the theory [1].

There are several reasons why it has been so hard to find a good theory in this field. Traditional metallic products can be defined by current computer-aided design (CAD) systems and manufactured by conventional processes; in fact, there are usually multiple ways to manufacture such traditional products. Therefore, such a product can be defined independent of the manufacturing process. But composite products are different because they are strongly dependent on the manufacturing processes used to produce them, and these manufacturing processes are subject to unceasing innovation and evolution. Since *the process defines the product* in these cases, traditional design theory and methodology (which hold that design can be decoupled from manufacturing) no longer apply. This is also the reason why traditional solid and geometrical modeling theories and methodologies implemented in current CAD systems are not sufficient to represent fibrous composite products. This problem is not restricted to composite products – there is a long list of other product categories for which current CAD systems do not provide adequate support [2].

Such deficiencies and drawbacks have not deterred engineers from using the current solid and geometrical modeling systems. They have augmented these systems with additional information, sometimes with ingenious informational artifacts, to carry out their mission. This has created a special breed of CAD systems that cater specifically to the composite product manufacturing domain. This is a bottom-up evolution, where pragmatic practical solutions are leading the development of engineering information systems and tools, without the benefit of a cohesive theory.

Such grass-roots developments have also resulted in major standardization efforts in ASME and ISO (International Organization for Standardization). Past and current editions of ASME Y14.37, ISO 10303-209, and ISO 10303-242 standards have codified the evolving industrial best practices in defining and modeling the information content of composite products [3-5]. These ASME and ISO standards are a combination of what the current solid and geometrical modeling systems can offer, and what additional information is required, to design and manufacture these composite products. What they lack is a comprehensive theory that explains, and expands upon, these sound industrial practices.

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay. "On Practice and Theory of Constructive Composite Geometry and Topology." Paper presented at ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, US. August 18, 2019 - August 21, 2019.

In a mathematical sense, composite products are cell complexes with geometrical and topological structures [2]. This paper synthesizes a theory of constructive composite geometry and topology from the industrial best practices codified in the ASME and ISO standards. It offers a procedural representation from which relevant information can be extracted and other representations can be computed. It bears some resemblance to the theory of constructive solid geometry [6, 7], but there are some significant differences due to the rich internal fibrous structures and the flexibility of constituents in composite products.

This paper is envisioned as a first attempt in postulating a theory that is close to practice for fibrous composite products, thereby initiating further discussion to strengthen the theory while still retaining strong relevance to practice. The major scientific and technical contributions of this paper include (1) the introduction of the concept of constructive composite geometry, with a tree representation for it, (2) the capture of some of the critical constructive composite topology information in the form of an adjacency graph, and (3) the use of conformal mapping of oriented surfaces in the ply lay-up. These notions systematize what is already practiced in industry and partially codified in recent standards from ASME and ISO.

The paper is organized as follows. Section 2 describes the constructive nature of composite product definition codified in the ASME standards. Then Section 3 offers a constructive composite geometry representation in the form of a hierarchical tree. Additional information is captured in Section 4 for constructive composite topology in the form of an adjacency graph. Section 5 addresses the geometry and topology of fibers embedded the plies. Section 6 concludes the paper with a summary and some directions for further research and standardization.

2 Constructive Nature of Composite Product Definition

This paper focuses on fibrous composite products, which are also known as fiber-reinforced plastics (FRP). In these composite products, thin fibers (of circular cross-section) made up of materials such as carbon, glass, or aramid (e.g., Kevlar) are embedded in a plastic resin medium called a matrix. Both fibers and resins play important roles in composite products. Fibers are the load bearing members and are the major contributors to the strength and stiffness. The resins transfer loads among the fibers and they provide much needed protection to the fibers from ambient environment (e.g., resistance to corrosion); they are also responsible for the ductility and toughness of the composite product. Thermosetting resins, which include epoxy and polyester, are currently very popular even though they are not recyclable. Thermoplastic resins are recyclable, but such resins are still under development for large-scale industrial use.

An important discrete module of a composite product is a *ply*, which is usually an arrangement (in potentially intricate geometric patterns) of reinforcement fibers in a resin matrix. The ply is often treated as a fabric; in fact, the technology and terminology of manufacturing a ply is strongly influenced by the textile industry. Fibers can be woven in special patterns called 'preforms,' which can be impregnated in resins to produce 'prepregs' [8, 9]. The ASME standard defines a ply as 'one discrete piece of manufactured material (e.g., fabric, tape, adhesive film)' [3]. Two or more plies can be stacked up to form a *laminate* (see Fig. 4 for an illustration).

Another important discrete module in a composite product is a *core*, which is a light weight component sandwiched between plies. The primary role of a core is to increase the 'section modulus' of thin-walled structures without increasing their weight considerably. A core can have hollow interiors that may be filled with air or special gases.

Composite products are produced by stacking up (or, laying up) plies, cores, and other items in a specified sequence, and then subjecting them to a curing process [8, 9]. Curing is a chemical process that enables polymers in the resins contained in the plies to cross-link, which produces a harder and more homogenous matrix within which the fiber reinforcements are firmly embedded. This leads naturally to a procedural representation, where the process for constructing a geometric and information model for the product mimics the manufacturing process that produces it. This notion is retained strongly in the current industrial practice for defining and modeling composite products.

This contrasts with how the geometry of many conventional metallic products are defined, say in a CAD system, using a set of geometrical construction commands. Such commands are intended to be better suited for the ease of constructing the *nominal* three-dimensional (3D) solid and geometric model of a part, ostensibly with little concern to the way in which the part is manufactured, say using traditional machining processes. This approach is strongly encouraged by other popular and long-standing ASME and ISO standards [10, 11] that enforce the dogma that product geometry definition should not be mixed with manufacturing process specification. The manufacturing concerns are, of course, taken into account during tolerancing these parts to ensure their manufacturability. While this good intention of separating design from manufacturing permits multiple ways to produce the same product – thereby enabling optimization on other metrics such as cost, quality, and time – this philosophy does not seem to apply to composite products.

The constructive nature of composite product definition has long been seized upon by engineers who came up with drafting and modeling schemes, while augmenting them with tables and specialized symbols. Such engineering schemes were initially standardized by ISO in 2001 for information exchange and later by ASME in 2012 for drawing practices [12, 13]. These standards have been updated recently in 2019 [3-5].

The standardization of composite product definition by ASME is best illustrated with an example. Figure 1 shows the plan view of a composite part. Following the 'dash number' convention, this part is denoted as -101 and it is constructed by stacking up several plies and a core as shown in a cross-sectional view in Fig. 2. It is clear from these two illustrations that this composite part is actually a bonded assembly of several plies and a core. In fact, it is common to refer to such a part as a 'bond assembly,' sometimes abbreviated as 'BOND ASSY.'

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay. "On Practice and Theory of Constructive Composite Geometry and Topology." Paper presented at ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, US. August 18, 2019 - August 21, 2019.

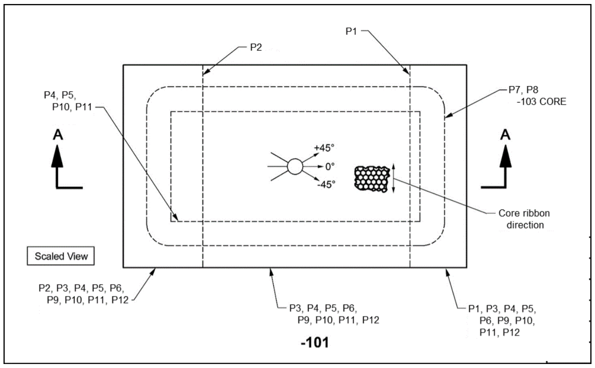


Fig. 1 Plan view of a composite part [3].

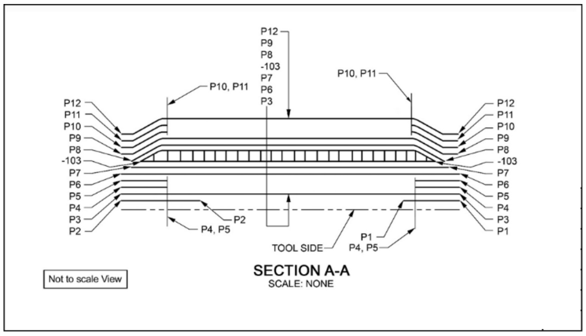
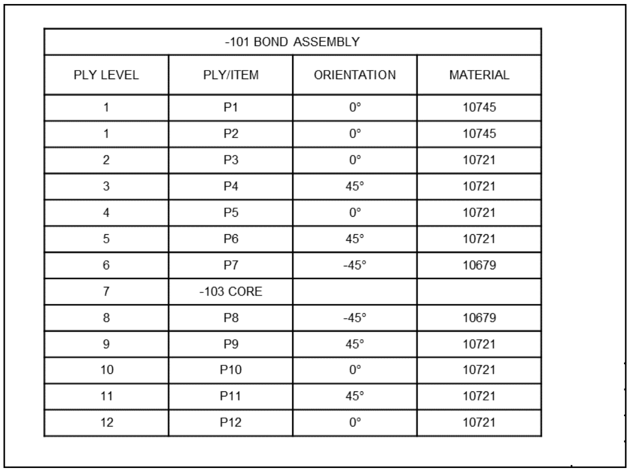


Fig. 2 Cross-sectional view of the composite part in Fig. 1 [3].

Table 1 Ply table for the composite part in Fig. 1 [3].



-101 BOND ASSEMBLY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	0°	10745
1	P2	0°	10745
2	P3	0°	10721
3	P4	45°	10721
4	P5	0°	10721
5	P6	45°	10721
6	P7	-45°	10679
7	-103 CORE		
8	P8	-45°	10679
9	P9	45°	10721
10	P10	0°	10721
11	P11	45°	10721
12	P12	0°	10721

The plan and (cross-sectional) elevation views in Figs. 1 and 2 are augmented by an important informational artifact called a *ply table* (also known as a *laminate table*) such as the one shown in Table 1. While these plan and elevation views follow the traditional drawing conventions for all industrial parts, the ply table is a special information that is associated only with composite products. In addition to the 'ply level' and 'ply/item' columns, the ply table in Table 1 has an 'orientation' column to indicate the fiber orientations that provide the necessary direction-dependent anisotropic properties. The last column in Table 1 identifies the material code for each ply, which can link to a much richer set of information specific to that ply. It is often useful to provide a 3D exploded view of the part as shown in Fig. 3 to accompany a ply table and sectional views.

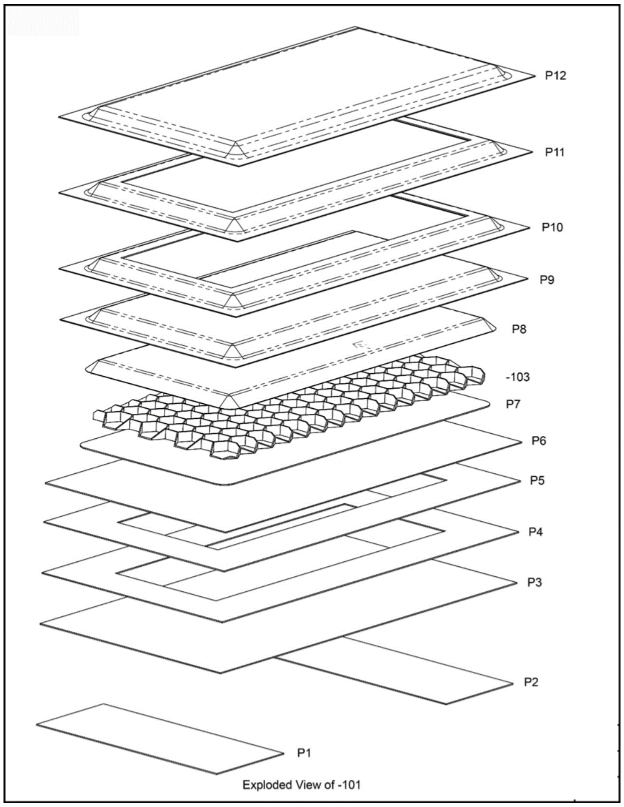


Fig. 3 A 3D exploded view of the composite part in Fig. 1 [3].

The figures and the ply table seen thus far provide a partial definition of a composite part in an *uncured* state. But this is not the final state of the product. Figure 4 explains how the ASME Y14.37 standard defines the uncured and cured states. In the uncured state on the left of Fig. 4, various plies are stacked up in a particular order to form a laminate. This laminate is then subjected to a curing process to produce the cured composite part on the right of Fig. 4. The ply interfaces disappear in this final product. It is interesting to note that much of the ASME and ISO standards are devoted to the definition and information modeling of the uncured state of the composite product and not to the 'net shape.' This important fact will be explored further in the rest of the paper.

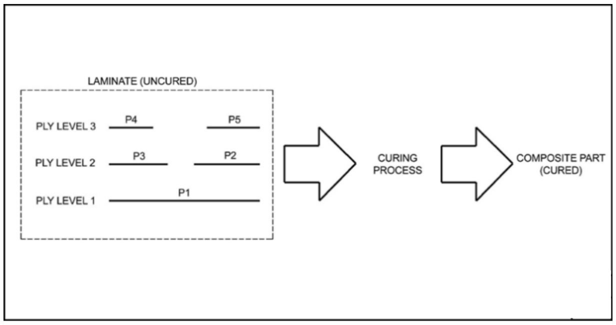


Fig. 4 Uncured and cured states in ASME standard [3].

The composites manufacturing process illustrated in these figures thus far may appear to be similar to semiconductor chip manufacturing. Both employ layered sequences of manufacturing processes, but there are important differences. One difference is that the semiconductor manufacturing uses subtractive processes (e.g., etching) as well as additive processes (e.g., sputtering). Another difference is that the composite manufacturing can produce complex 3D parts. Nevertheless, some of the ideas on geometrical and information modeling associated with semiconductor products may be useful for defining composite products.

While the logical design of a semiconductor chip can be carried out without invoking the manufacturing processes, the physical design of such a chip is strongly dictated by the sequence of process steps to which a semiconductor wafer is subjected. This notion will be carried further for composite products in rest of the paper. Section 3 will address the geometrical aspect, while Section 4 will consider the topological aspect of the constructive nature of composite product definition.

3 Constructive Composite Geometry

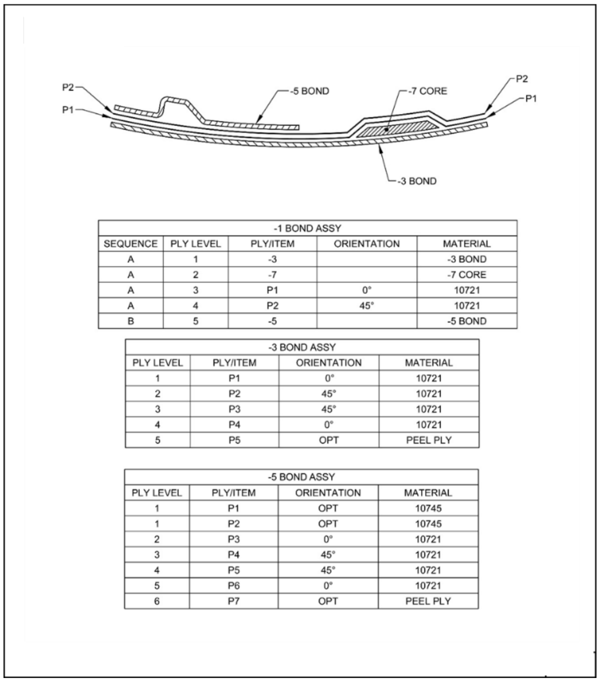
The example illustrated in Section 2 brings out some salient features of plies and ply tables in composite product definition. These include the following:

1. A ply is used both as a discrete physical artifact and as an information container. As a physical artifact, it contains literally the fibers and resins. As an information container it encapsulates complex information about the geometry, topology, and material of the fibers and resins in each instance of the ply. This seems to be the only practical way to handle the physical and information complexity of a composite product.

2. A ply table provides an ordered sequence (arranged from the top row to the bottom row) of the placement of the plies and cores. In a physical sense, this is part of the recipe for the manufacturing process that leads up to the uncured state of the product. In an informational context, this implies some geometrical and topological adjacency of not only the plies but also the fibers contained in them. Since the fibers remain firmly embedded in the hardened plastics after curing, this adjacency information of the plies is inherited by the fibers contained in the plies, both before and especially after curing. In other words, the ply table captures some crucial adjacency relationships among the fibers and these relationships remain invariant under the curing operation.

3. A ply table alone, of course, will not provide all the geometrical and topological information. The accompanying cross-sectional views (such as Fig. 2) and 3D exploded views (such as Fig. 3) supply much needed details. From a geometrical modeling perspective, it is interesting to note that a ply is depicted as a curve in a cross-sectional view in Fig. 2 and as a surface in a 3D exploded view in Fig. 3, even though a ply is a three-dimensional object with some definite thickness. This type of abstraction has a deeper implication than being just a convenient way for graphical presentation.

Further insights from the use of ply tables can be gained from examples shown in Figs. 5 and 6 taken from the ASME standard [3].



-1 BOND ASSY				
SEQUENCE	PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
A	1	-3		-3 BOND
A	2	-7		-7 CORE
A	3	P1	0°	10721
A	4	P2	45°	10721
B	5	-5		-5 BOND

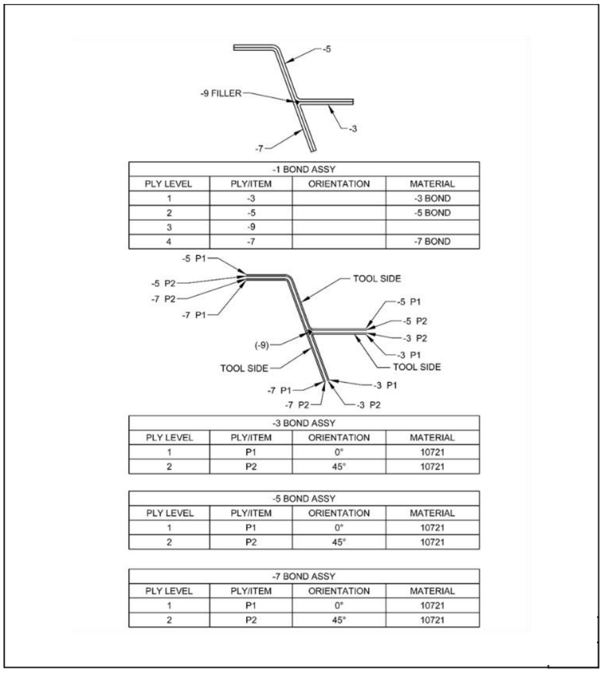
-3 BOND ASSY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	0°	10721
2	P2	45°	10721
3	P3	45°	10721
4	P4	0°	10721
5	P5	OPT	PEEL PLY

-5 BOND ASSY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	OPT	10745
1	P2	OPT	10745
2	P3	0°	10721
3	P4	45°	10721
4	P5	45°	10721
5	P6	0°	10721
6	P7	OPT	PEEL PLY

Fig. 5 A multi-stage bonded assembly [3].

Figure 5 shows a cross-sectional view and a set of ply tables for a multi-stage bonded assembly, which means that sub-assemblies such as '-3 BOND ASSY' and '-5 BOND ASSY' can be precured and then be bonded into the '-1 BOND ASSY.' It is interesting to observe that the ply table for '-1 BOND ASSY' contains links to other ply tables for '-3 BOND ASSY' and '-5 BOND ASSY,' thus suggesting a hierarchical tree structure for organizing this type of information. A similar example for a complex bonded assembly is shown in Fig. 6, where one ply table contains links to other ply tables.

In addition to the plies, other items such as '-7 CORE' in Fig. 5 and '-9 FILLER' in Fig. 6 can be included in a ply table. These cores and fillers can be modeled as 3D solids; even the precured subassemblies (such as '-3 BOND ASSY' and '-5 BOND ASSY' in Fig. 5; and '-3 BOND ASSY,' '-5 BOND ASSY,' and '-7 BOND ASSY' in Fig. 6) can be modeled as 3D solids. Representations of these solids then become both physical and information containers for other objects, just as the plies discussed earlier.



-1 BOND ASSY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	-3		-3 BOND
2	-5		-5 BOND
3	-9		
4	-7		-7 BOND

-3 BOND ASSY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	0°	10721
2	P2	45°	10721

-5 BOND ASSY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	0°	10721
2	P2	45°	10721

-7 BOND ASSY			
PLY LEVEL	PLY/ITEM	ORIENTATION	MATERIAL
1	P1	0°	10721
2	P2	45°	10721

Fig. 6 A complex bonded assembly [3].

The notion of ply tables and their engineering use have by now been well established and deeply entrenched in composite manufacturing industry. In fact, this practice is so stable that all the earlier and current versions of the ASME and ISO standards have codified them and expanded them to cover other manufacturing processes such as braiding and pultrusion [8, 9]. So, it is reasonable to propose a constructive composite geometry (CCG) representation in the form of a hierarchical tree that is equivalent to a ply table.

Figure 7 shows a CCG tree for the '-101 BOND ASSEMBLY' associated with the ply table in Table 1 and the exploded view in Fig.3. The root node in the tree is the composite product denoted as -101 BOND ASSEMBLY. The leaf nodes are the plies, the core, and the tool. More information, such as 'orientation' and 'material' found in Table 1, can be associated with the leaf nodes as entities and attributes. The interior nodes in the CCG tree in Fig. 7 are denoted by the symbol ⊕ and it stands for an 'adjoin' operation.

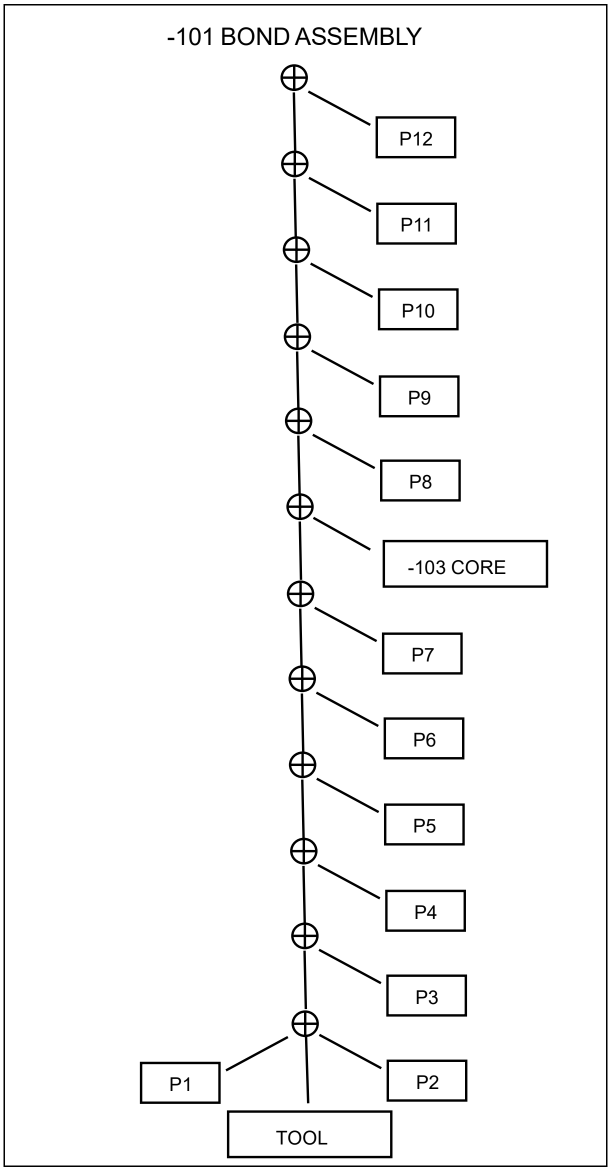


Fig. 7 A CCG tree for the -101 BOND ASSEMBLY in Table 1 and Fig. 2.

The adjoin operation, denoted by the symbol ⊕, is equivalent to a physical bonding (or gluing) of various plies and other items in the *uncured* state in a laminate (see Fig. 4). After curing, the thermosets or thermoplastics undergo chemical bonding, thereby rendering the adjoin operation as an approximation to a ‘set union’ operation among these plies in the *cured* state. Thus, the interpretation of the adjoin operation is state-dependent.

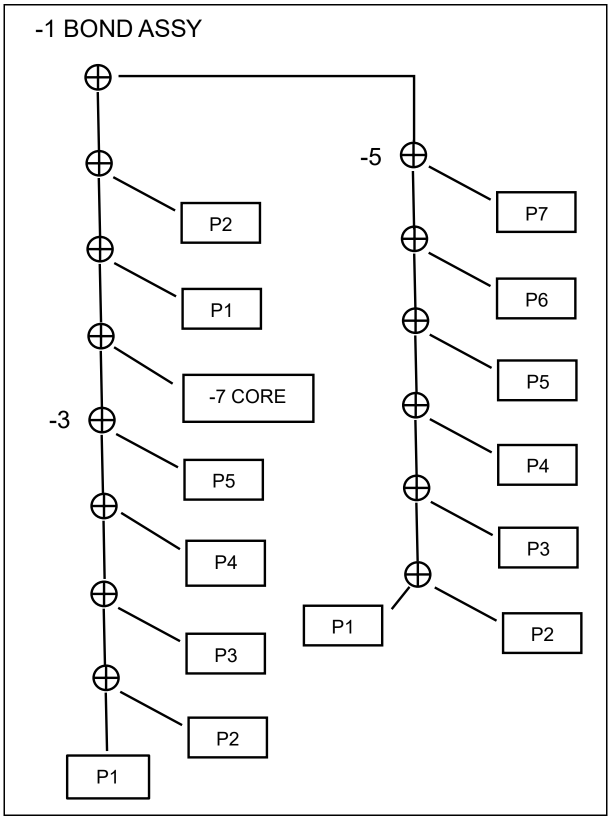


Fig. 8 A CCG tree for the -1 BOND ASSY in Fig. 5.

As a further example, Fig. 8 shows the CCG tree associated with the ply tables and figure in Fig. 5. At the root is the multi-stage bonded assembly denoted as ‘-1 BOND ASSY.’ Similarly, Fig. 9 shows the CCG tree associated with the ply tables and figures in Fig. 6, with the root node denoted as ‘-1 BOND ASSY.’ The hierarchical tree structures seen in Figs. 7, 8, and 9 are a direct consequence of the ply table structures standardized and shown in Table 1 and Figs. 5 and 6.

It is instructive to examine the three CCG trees in Figs. 7, 8, and 9 in some detail, along with their corresponding ply tables and figures.

1. *Tool*: Ply lay-up starts with a tool surface on a tool. The ‘tool side’ surface (often denoted as in Figs. 2 and 6) plays the role of a ‘datum feature’ in the composite product definition; a physical tool (with a surface on which the plies are laid) serves as the manufacturing/assembly fixture for the ply lay-up. Using this tool surface as a datum simulator, proper datums and datum systems can be established for the composite product definition [10, 11]. 3D modeling of the tool and the tool surface can be accomplished using currently available solid and geometrical modeling systems.

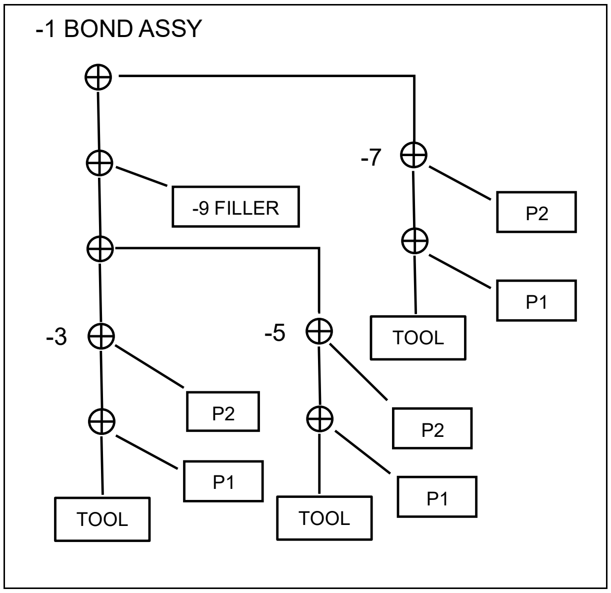


Fig. 9 A CCG tree for the -1 BOND ASSY in Fig. 6.

2. *Core/Filler*: Items such as core and filler can be modeled as 3D solids. Some of the cores, such as the one with honeycomb structure shown in the exploded view of Fig. 3, can have interior cavities that are filled with air or special gases. Figure 1 shows the ‘core ribbon direction’ for proper orientation of the core with respect to a datum reference frame, which can be established using the ‘tool side’ information mentioned earlier.

3. *Adjoin*: The adjoin operation, denoted by the symbol ⊕, is a binary operation like ‘addition’ and ‘union.’ It can be viewed as a simple gluing operation preserving a physical bonding in an uncured state; the interface between the adjoined objects are preserved in this state. In a cured state, the interfaces between the plies disappear due to chemical bonding and this is similar to the result of a ‘set union’ operation.

   The algebra of the adjoin operation requires some careful consideration. It is tempting to assign commutativity (that is, $P1 \oplus P2 = P2 \oplus P1$) and associativity (that is, $P1 \oplus (P2 \oplus P3) = (P1 \oplus P2) \oplus P3$) to this operation. But, it is not advisable to do so, due to the strong order dependency in the ply lay-up as described below.

4. *Ply*: As mentioned earlier, ASME defines a ply as ‘one discrete piece of manufactured material (e.g., fabric, tape,

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay. "On Practice and Theory of Constructive Composite Geometry and Topology." Paper presented at ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, US. August 18, 2019 - August 21, 2019.

adhesive film)' [3]. The primary geometrical representation of a ply is that of a surface with certain thickness. Using the textile metaphor, a ply is like a cloth or a fabric. It can be viewed as a flat surface before a lay-up, and it can be draped as a curved surface on what has been already laid-up or on a tool surface.

The ply is flexible, and it can be tucked and squeezed to conform to the surface to which it is adjoined. The ply thickness may not remain constant during this lay-up operation; the main objective is to make the 'bottom side' of the ply stick to the surfaces on which it is laid without leaving any voids or holes. This poses some interesting challenges in modeling a ply as a 3D solid. This also contributes to the caution about commutativity and associativity of the adjoin operation raised earlier.

5. *Orientation*: Plies also encapsulate reinforcement fibers, both physically and informationally. This will be addressed later in Section 5 in some detail. An important information related to the fibers in a ply is their relative orientation with respect to an external reference frame, which can be established using datums mentioned earlier.

   The angles mentioned in the 'orientation' column of the ply tables seen thus far provide such information. Additionally, a local reference frame called a 'rosette' (as shown in the center of Fig. 1) can be affixed on a ply surface – sometimes in many places on a ply surface – to specify the orientation of the fibers within a ply. Figure 10 shows several examples of such ply orientation symbols associated with rosettes. An example of the use of a guide curve and rosettes is illustrated in Fig. 11. More rosette types and their usage can be found in the recent ASME standard [3]. A rosette is like a two-dimensional compass placed on an undulating terrain, indicating the local directions along which the fibers in a ply should be oriented.

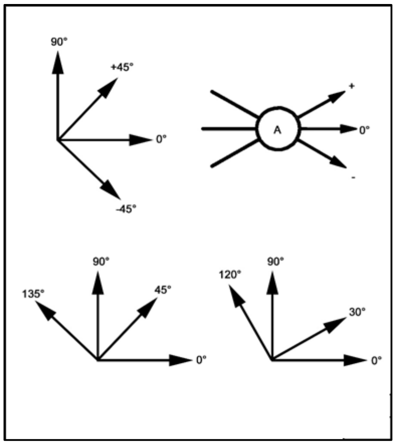


Fig. 10 Examples of ply orientation symbols [3].

6. *Material*: The identifier mentioned in the 'material' column of the ply tables provides the link to much richer information to various other material information about the plies, including critical information about the reinforcement fibers and resins.

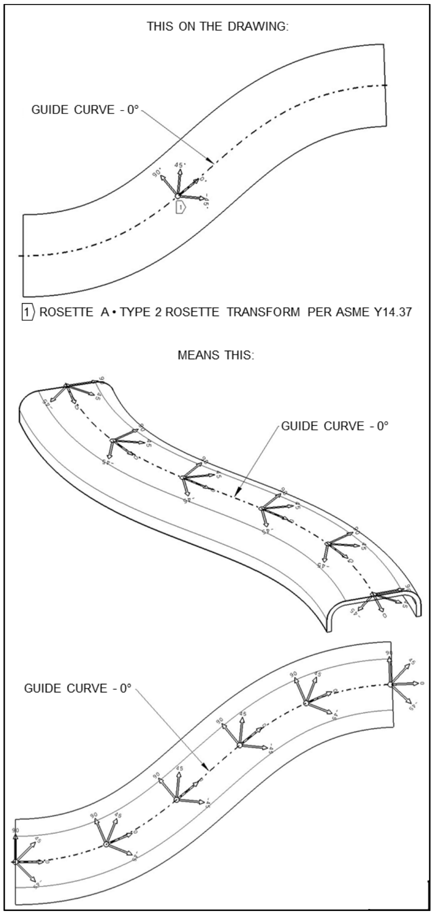


Fig. 11 An example of using a particular type of rosette [3].

The CCG tree is similar to the more familiar constructive solid geometry (CSG) tree [6, 7]. Both are procedural representations from which other representations can be derived. For example, a boundary representation (also known as B-rep) of a solid can be derived from its CSG representation to facilitate several important applications. Following this line of thinking, one may expect that an explicit cell-complex representation of a composite product may be derived from its CCG representation. It turns out that this requires a careful consideration of the topology of the composite products, which will be addressed next.

4 Constructive Composite Topology

The concept of topology captures the notion of adjacency of geometrical objects without paying too much attention to the underlying geometrical details. This does not mean that the geometrical details do not matter; the adjacency of 3D cells in a composite product critically depends on how various plies and other items are positioned in an assembly. Topology, represented as an adjacency graph, turns out to be one of the most robust properties that should be captured in a composite product definition and controlled in its manufacturing.

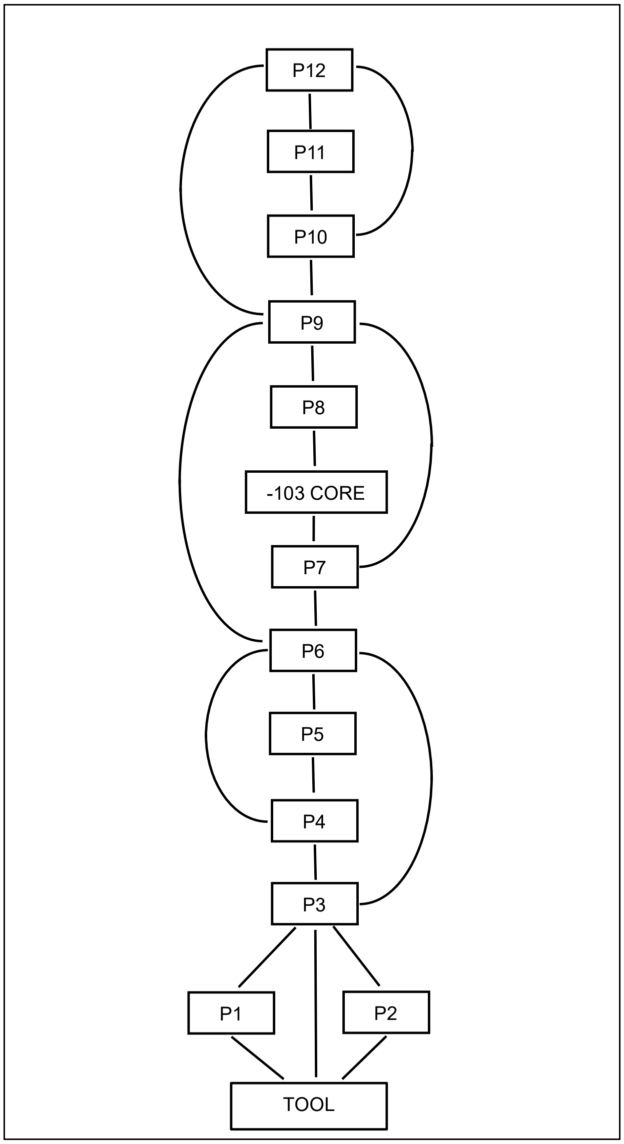


Fig. 12 A CCT graph for the -101 BOND ASSEMBLY in Table 1 and Figs. 2 and 7.

To establish the adjacency of 3D objects in a composite product, consider the plies, cores, and fillers as 3D solids. Figure 12 shows an adjacency graph of the constructive composite topology (CCT) for the -101 BOND ASSEMBLY in Table 1 and Figs. 2 and 7. A similar CCT graph for the -1 BOND ASSY seen earlier in Figs. 5 and 8 is shown in Fig. 13. Also, the CCT graph for the -1 BOND ASSY encountered earlier in Figs. 6 and 9 is shown in Fig. 14. In all these three CCT graphs, the nodes are the plies, cores, fillers, tools, and precured subassemblies; these are displayed as rectangular boxes. The arcs represent the fact that the solids in the nodes joined by each arc have a finite surface area of contact in the uncured state of the bonded assembly.

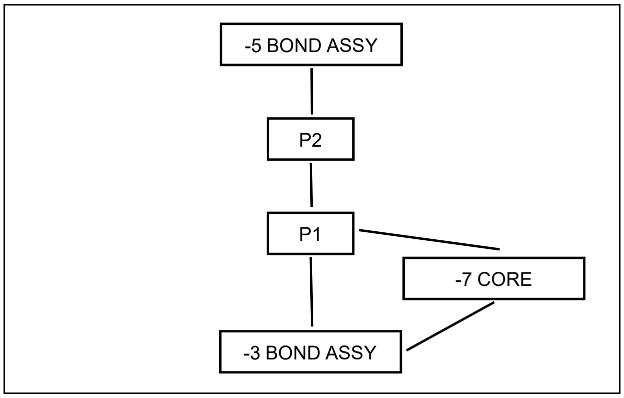


Fig. 13 A CCT graph for the -1 BOND ASSY in Figs. 5 and 8.

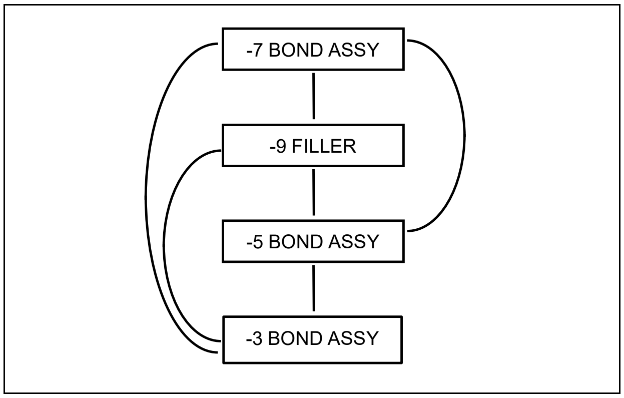


Fig. 14 A CCT graph for the -1 BOND ASSY in Figs. 6 and 9.

Figure 15 illustrates the cell structure and the adjacency relationship using an example previously encountered in Figs. 5, 8, and 13. The cross-sectional view in Fig. 15 captures only a small portion of the bonded assembly, and the dimensions are exaggerated for clarity. The core and the plies in this example are bonded to the objects on which they are laid. It is important that no 'air bubble' or 'air hole' is introduced in the lay-up process or

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay. "On Practice and Theory of Constructive Composite Geometry and Topology." Paper presented at ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, US. August 18, 2019 - August 21, 2019.

in the 3D model creation. This poses a challenging problem when approximations to the 3D models, such as tessellated polyhedra, are used for 3D representations of various constituents in a bonded assembly.

Each of the four objects in Fig. 15 is considered as a 3D cell. If any two of these 3D cells have a contact over a surface area, then they are joined by an arc in the CCT graph shown in Fig. 13. This illustrates the concept of adjacency used in all the three CCT graphs shown in Figs. 12, 13, and 14.

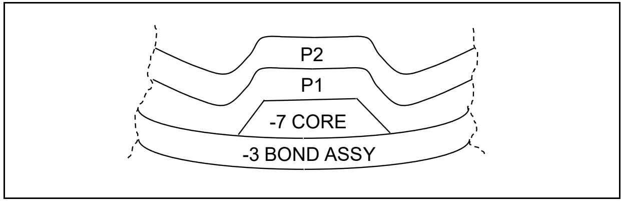


Fig. 15 Cross-section of a portion of an approximate 3D model of the cell-complex of the '-1 BOND ASSY' in Figs. 5 and 8. Drawing is not to scale.

As a further illustration of the CCT graph, consider the example shown in Figs. 1, 2, and 3. Figure 16 shows the cell structure and the adjacency relationship using a cross-sectional view of an approximate 3D model. Such topological information is represented only in a CCT graph; it is not available explicitly either in an exploded view such as Fig. 3 or in a ply table such as Table 1.

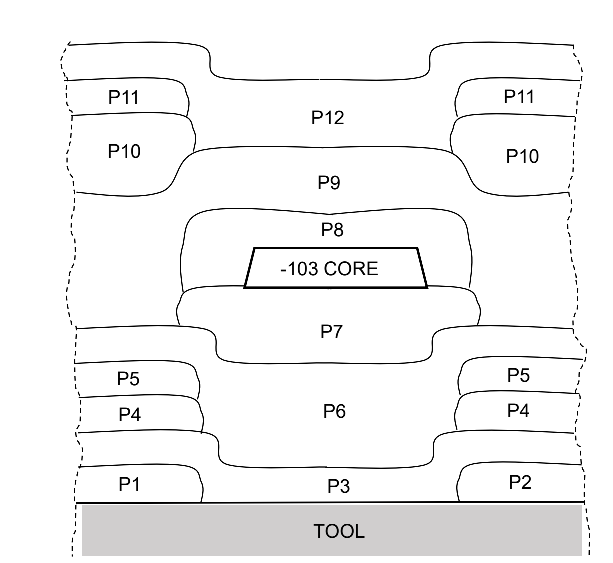


Fig. 16 Cross-section of a portion of an approximate 3D model of the cell-complex of the '-101 BOND ASSEMBLY' in Figs. 1, 2, and 3. Drawing is not to scale.

It is also clear from Figs. 15 and 16 that 'ply thickness' is not a simple number after the ply is stacked up in a laminate in an uncured state. These plies are indeed pliable, and they undergo a homeomorphic transformation as they go from the state of a flat fabric into a stacked-up ply in a laminate in an uncured state. Again, it is important to avoid any air pockets or air holes (neither 'through holes' nor 'blind holes') during the stack-up process.

With these examples as preliminaries, the following observations can be made about the structure and property of the CCT graph.

1. *The CCG tree provides a useful input to build the CCT graph*. The adjoin operation in the CCG tree suggests an adjacency relationship between its operands. CCT graph contains these relationships, but it also captures more adjacency relationships.

2. *The CCT graph is a subgraph of the dual graph of the cell complex*. Algebraic topology of a cell complex, as applied to a composite product, consists of a hierarchy of 3-cells (solids) bounded by 2-cells (surfaces) bounded by 1-cells (edges) bounded by 0-cells (vertices) [14]. A dual graph of this cell complex will include cells of all these dimensions; but the CCT graph captures only the 0 and 1 dimensional subsets of this dual graph.

3. *The plies undergo a homeomorphic transformation in the lay-up process*. The plies start out as flat fabrics. As a ply is draped in the lay-up process by tucking, stretching, and squeezing, it undergoes a homeomorphic transformation, which is a basic notion in topology [14]. Under such a homeomorphic transformation, neighboring points of a ply remain as neighbors. This is an important property that is observed during the lay-up process to ensure the integrity of the reinforcement fibers (e.g., no tearing) contained in the plies. It is also an important property that can be used in 3D modeling of the laid-up plies in an uncured state.

   Such homeomorphic transformations are shown, albeit with some exaggeration, in Figs. 15 and 16. The ply thickness can undergo changes under this transformation. But it can be argued that the volume of the ply remains the same, obeying the conservation of mass (before curing). Thus, the most robust statement that can be made is that the plies undergo a volume-preserving homeomorphic transformation during the lay-up process.

4. *The CCG tree and CCT graph serve as important primary elements of the procedural representation*. The robustness of CCG trees and CCT graphs has been well established in the industrial practice though decades of use in the form of ply tables and accompanying figures. Taking them as important primary representations, explicit 3D representations may be derived as needed for the uncured laminates and cured bonded assemblies. This is similar to the derivation of a B-rep from a CSG representation of a

solid, where the CSG is the primary (procedural) representation and B-rep is the derived (explicit) representation [6, 7]. Specific composite manufacturing properties, such as volume-preserving homeomorphic transformations mentioned above, may be utilized for constructing the 3D representations of such cells.

5 Geometry and Topology of Reinforcement Fibers

Plies contain reinforcement fibers that contribute much of the strength and stiffness to the composite products. These fibers are combined using textile technologies, such as weaving, stitching, seaming, and braiding, to form various patterns in a fabric. Such fabrics can be created as ‘preforms’ as shown in Fig. 17, where each colored strip contains many fibers. These preforms can also be impregnated in resins to create plies as ‘prepregs’ that can be handled as cloths, as shown in Fig. 18.

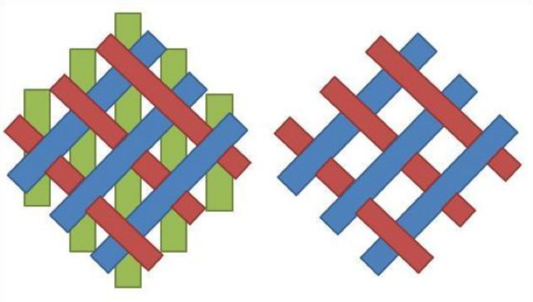


Fig. 17 Examples of preforms in creating plies [15].

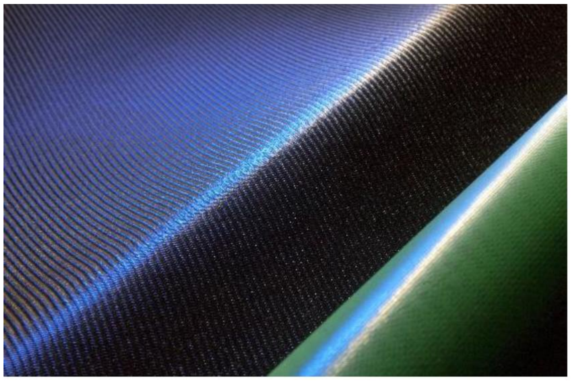


Fig. 18 Example of prepregs in creating plies [16].

The ASME and ISO standards do not address the details of the fiber arrangements in plies. Instead, they refer to other standards such as the ASTM standard on stiches and seams [17]. The ‘material’ attribute in the ply table provides links to such fiber arrangement details within each ply.

The geometry and topology of the fiber arrangements in a ply can benefit from research literature on textiles [18]. When a ply is draped as a cloth to form a laminate, the local orientations of the fibers receive a great deal of attention in practice, as illustrated in Fig. 11 using rosettes. From such standardized practices, a theory for the geometry and topology of bundles of reinforcement fibers can be derived using the following two important invariance properties:

1. *Conformal mapping to preserve local fiber orientations*. When a flat fabric containing reinforcement fibers is draped onto a curved surface, as shown in Fig. 11 for example, the local orientations of fibers are preserved within each ply. This practice can be directly related to the theory of conformal mapping of oriented surfaces, which preserves the angles and shapes of infinitesimally small figures but not necessarily their sizes or curvatures [19]. In fact, the standardized definitions of rosettes [3] and their usage in practice can be formalized mathematically as conformal mapping of oriented surfaces. Such mappings occur in the lay-up of each ply to form a laminate before the curing operation. These fiber orientations are also expected to be preserved in the cured state of the composite product, within some orientation angle tolerance.

   In a typical conformal mapping of oriented surfaces, the ‘domain’ surface is parameterized using (u, v) coordinates, and it is mapped to a ‘range’ surface that is parameterized using (u', v') coordinates. It is then customary to talk about u-curves being mapped to u'-curves, and v-curves being mapped to v'-curves. In the context of manufacturing of composite products, these u-curves, u'-curves, v-curves, and v'-curves take on a physical meaning. These curves are the representations of the fibers in a semi-discrete form. An interesting constraint that arises in such conformal mapping is that the fiber lengths are preserved in the lay-up process.

2. *Preservation of fiber adjacency between layers*. As noted in Fig. 4, the cured composite part does not preserve the interfaces between the plies. However, the adjacency relationships among the plies and other items established in the CCT graph (such as Figs. 12, 13, and 14) are inherited by the bundles of fibers contained in these plies. This provides two important types of information about the bundles of fibers in a cured composite product: (1) Each bundle of fiber is given an identifier that is inherited from the ply within which it is impregnated; (2) A partial spatial ordering of these bundles of fibers in the cured composite product can be derived from the CCT graph. Such information is important for inspection (e.g., nondestructive testing) and structural analysis purposes. For example, a cluster analysis of fibers in a 3D computed tomography image of a cured composite product may benefit greatly from a prior knowledge of the adjacency of bundles of the fibers.

6 Summary and Directions for Future Research

This paper took some initial steps towards a theory of constructive composite geometry and topology. This theory is synthesized from industrial best practices codified in recent ASME and ISO standards [3, 4]. It is also motivated by the need for better scientific tools to explore new frontiers in modeling material structures that are produced by modern manufacturing technologies [2]. The major features of this theory are:

1. A CCG tree, which maintains a one-to-one mapping with the ply/laminate table standardized by ASME and ISO. An 'adjoin' operation in the CCG tree defines the bonding (both physical and chemical) between various plies and other items in the assembly.

2. A CCT graph, which captures the adjacency relationship among plies and other items in an uncured state of a laminate, and this adjacency relationship is inherited by the bundles of fibers in the cured composite product.

3. A conformal mapping between a flat ply surface and a draped ply surface in a laminate; this mapping preserves the local orientations of fibers (in terms of angle between fibers) as standardized by ASME and ISO using rosettes and other orientation specifications.

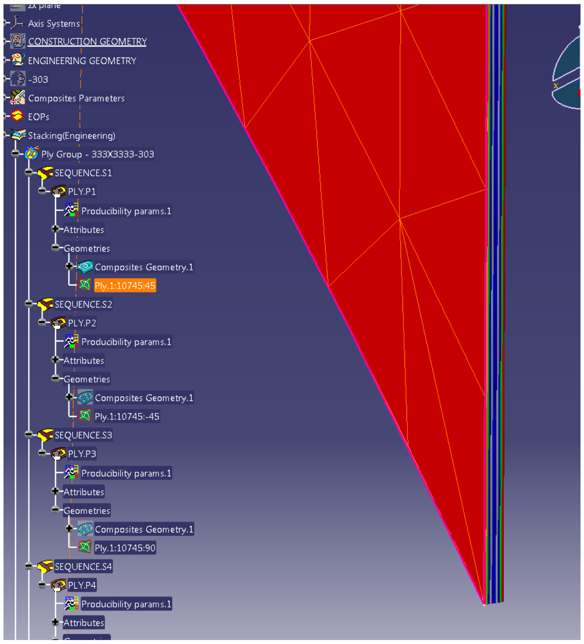


Fig. 19 Example of a composite part used for testing standardized 3D model data exchange [20].

Some of these theoretical abstractions have been implemented already in practice. In major CAD systems that support composite products, a 'model tree' is a prominent visual component of the user interface; this can be easily related to the CCG tree. Figure 19 shows a screen shot of a composite part used for testing standardized 3D model exchange [20]. It clearly shows a tree-like representation for the ply stack-up under the 'Stacking (Engineering)' node in the model tree displayed on the left.

While these are encouraging signs, more research and standardization need to be undertaken along the following lines:

- The CCG tree and CCT graph can be extended to cover more composite manufacturing technologies such as infusion, pultrusion, and braiding. Recent ASME and ISO standards have already taken initial steps in this direction [3, 4].

- 3D modeling of uncured and cured states of composite products can be improved considerably, with proper attention to the conformal mapping of ply surfaces and the homeomorphic transformation of ply (and other) solid cells, as described in the body of the paper.

- Better harmonization of ASME and ISO standards is needed so that the composite product definition practices and 3D models for information exchange are properly aligned.

Acknowledgments and a Disclaimer

The authors gratefully acknowledge the help and support of ASME and ISO standards committee members. Any mention of commercial products in this paper is for information only; it does not imply recommendation or endorsement by NIST.

References

[1] Grandine, T., 2013, "Geometric Interoperability for Composite Manufacturing," Panel on Geometric Interoperability, ASME Man. Sci. Eng. Conf., Madison, WI.

[2] Regli, W., Rossignac, J., Shapiro, V., and Srinivasan, V., 2016, "The New Frontiers in Computational Modeling of Material Structures," Com. Aided Des., **77**, pp. 73-85.

[3] ASME Y14.37-2019, 2019, *Product Definition for Composite Parts*, The American Society of Mechanical Engineers, New York.

[4] ISO 10303-242:2019, 2019, *Industrial automation systems and integration – Product data representation and exchange –Part 242: Application protocol: Managed model-based 3Dengineering*, International Organization for Standardization, Geneva, Switzerland.

[5] Hedberg Jr, T.D., Feeney, A.B., and Srinivasan, V., 2019, "An Analysis of Recent Standards on Composite Product Models to Enable Digital Transformation of Composite Product Manufacturing," MSEC2019-2783, ASME Man. Sci. Eng. Conf., Erie, PA.

[6] Hoffmann, C.M., 1989, *Geometric and Solid Modeling: An Introduction*, Morgan Kaufmann, Palo Alto, CA.

[7] Srinivasan, V., 2004, *Theory of Dimensioning: An Introduction to Parameterizing Geometric Models*, Marcel Dekker, NY.

[8] Campbell, F.C., 2004, *Manufacturing Processes for Advanced Composites*, Elsevier, Oxford, U.K.

[9] Strong, A.B., 2008, *Fundamentals of Composites Manufacturing*, 2nd Edition, Society of Manufacturing Engineers, Dearborn, MI.

[10] ASME Y14.5-2009, 2009, *Dimensioning and Tolerancing*, The American Society of Mechanical Engineers, New York.

[11] ISO 1101:2017, 2017, *Geometrical Product Specifications (GPS) – Geometrical Tolerancing – Tolerances of Form, Orientation, Location and Run-out*, International Organization for Standardization, Geneva, Switzerland.

[12] ISO 10303-209:2001, 2001, *Industrial Automation Systems and Integration – Product Data Representation and Exchange – Part 209: Application Protocol: Composite and Metallic Structural Analysis and Related Design*, International Organization for Standardization, Geneva, Switzerland.

[13] ASME Y14.37-2012, 2012, *Composite Part Drawings*, The American Society of Mechanical Engineers, New York.

[14] Hatcher, A., 2002, *Algebraic Topology*, Cambridge University Press, U.K.

[15] Wphallig. "Braids." Digital image. *Wikimedia Commons*. November 26, 2013. Accessed October 21, 2018. Licensed under CC BY-SA 3.0. https://commons.wikimedia.org/w/index.php?curid=29864409.

[16] Cross, N., 2009. "Gurit - a World Leader in Epoxy Prepreg Technology." Digital image. *Flickr*, January 30, 2009. Accessed October 25, 2018. Licensed under CC BY-ND 2.0. www.flickr.com/photos/80188450@N03/8138514559.

[17] ASTM D6193-16, 2016, *Standard Practice for Stiches and Seams*, ASTM International, West Conshohocken, PA.

[18] Grunbaum, B., and Shephard, G.C., 1980, "Satins and Twills: An Introduction to the Geometry of Fabrics," Mathematics Magazine, **53**(3), pp. 139-161.

[19] Kreyszig, E., 1991, *Differential Geometry*, Dover, New York.

[20] https://www.cax-if.org/production_model.html. Accessed Jan. 23, 2019.

Hedberg Jr., Thomas; Barnard Feeney, Allison; Srinivasan, Vijay. "On Practice and Theory of Constructive Composite Geometry and Topology." Paper presented at ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Anaheim, CA, US. August 18, 2019 - August 21, 2019.

On Selecting Channel Parameters for Public Safety Network Applications in LTE D2D Communications

Siyuan Feng, Hyeong-Ah Choi
The George Washington University
Washington, District of Columbia 20052
Email: {ff910829, hchoi}@gwu.edu

David Griffith and Richard Rouil
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
Email: {david.griffith, richard.rouil}@nist.gov

***Abstract*— The Third Generation Partnership Project (3GPP) defines various pre-configured channel parameters for the Long-Term Evolution (LTE) Device-to-Device (D2D) communications with Physical Sidelink (SL) Channels. In this paper, we investigate the impacts of channel parameter settings on the performance of content deliveries for Public Safety Network (PSN) applications in Out-of-Coverage (OOC) scenario. We first measure the reliability of the SL channels under various sets of channel parameters using Monte Carlo simulations. Then, for a given PSN application, the acquired reliability results are utilized to help determining the amount of delay that is to be introduced to the system, such that the throughput requirement for the application is assured during the transmissions. To the best of our knowledge, this is the first LTE D2D work that focuses on OOC mission-critical communications performance in group traffic settings. Our results are valuable to both network operators, for using them as references in selecting a best set of channel parameters, and to future studies on more complex transmission patterns and network scenarios using D2D communications in PSNs.**

I. Introduction

A. Background

During disasters, network infrastructures become susceptible to damages, and sometimes the entire network connection could be cut out [1]. To resolve this challenge we investigate the mission-critical communication (MCC) performances of the LTE Device-to-Device (D2D) communication protocol, ProSe (Proximity Services), when operating in Out-of-Coverage (OOC) environment under communication Mode 2.

Mode 2 communications in ProSe utilize the LTE Uplink resources to form Sidelink (SL) communication channels known as the ***P**hysical **S**idelink **C**ontrol **Ch**annel* (PSCCH) and the ***P**hysical **S**idelink **S**hared **Ch**annel* (PSSCH) that allow User Equipment (UEs) to establish direct communications, broadcasting in nature, without any backhauling. Due to this reason, it is considered as a prominent technology for group voice/data deliveries in OOC scenarios, both in the current LTE PSN settings and for ubiquitous applications in the future Fifth Generation New Radio (5G NR) settings.

However, since in OOC scenario there is no centralized coordination of the UEs, individual UEs allocate their own resources in the SL Resource Pool (RP) autonomously. This competitiveness among devices can cause severe packet losses. In this work, we adopt the packet loss models proposed in [2], [3], where packet loss is caused by Collision and Half-Duplex Effect (HDE). Later, we review the packet loss model and how to detect packet loss in the channels in Section III.

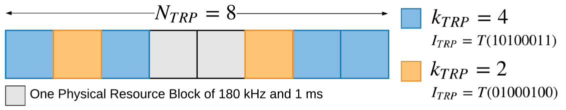


Fig. 1. **Time Resource Pattern (TRP)**: I_{TRP} is the index of the TRP mask which is a binary bit map of length N_{TRP}; when a mask bit is 1 then that bit is to be used for transmitting one data TB in the PSSCH; k_{TRP} is the number of subframes a UE can allocate its TBs within one N_{TRP} length, hence representing the number of 1's in a TRP mask. We fix $N_{TRP} = 8$.

In addition, due to its broadcasting nature, D2D communication does not have acknowledgment. To mitigate potential packet losses, two methods are specified by 3GPP [4]: $i)$ by transmitting duplicate copies of the Transport Blocks (TBs); specifically, two copies for the Sidelink Control Information (SCI) TBs, and four copies for the associated data message TBs. As well as, $ii)$ by limiting, k_{TRP}, the number of subframes each UE can allocate in every $N_{TRP} = 8$ subframes in the PSSCH, where TRP stands for Time Resource Pattern. The concept of TRP is illustrated in Fig. 1 with more details.

B. Problem Description

To test the performance of the proposed D2D solution in mission-critical settings, we considered an OOC group communication scenario with Mode 2 communications among N_u Half-Duplex UEs held by First Responders (FRs), all within each others' proximity. Now assume the group leader, denoted by UE-1, has an MCC message that must be disseminated to ALL the other $(N_u - 1)$ UEs in the group.

§ First, we examine the successful transmission probabilities of the SL channels by defining the **PSCCH Reliability** $\mathbb{P}\{\mathcal{S}_C\}$ and the **PSSCH Reliability** $\mathbb{P}\{\mathcal{S}_D\}$ as the probabilities of the following two events, respectively:

• $\mathcal{S}_C$: **successful reception of leader's SCI TB(s)**, "the event where all the other $(N_u - 1)$ UEs would receive at least one copy of UE-1's non-collided SCI TB in the PSCCH of a single independent SL period."

• $\mathcal{S}_D$: **successful reception of leader's data TB(s)**, "the event where all the other $(N_u - 1)$ UEs would receive at least one copy of UE-1's non-collided data TB in the PSSCH of a single independent SL period where event $\mathcal{S}_C$ took place."

Feng, Siyuan; Choi, Hyong-Ah; Griffith, David W.; Rouil, Richard A. "On Selecting Channel Parameters for Public Safety Network Applications in LTE Direct." Paper presented at IEEE Consumer Communications & Networking Conference (CCNC), Las Vegas, NV. January 10, 2020 - January 13, 2020.

§ Secondly, we consider a PSN application with a certain throughput requirement of $\mathcal{Q}$ bits/sec that is to be fulfilled, i.e., the Guaranteed Bit Rate (GBR) of that application. We are interested in the question that: while guaranteeing event $\mathcal{S}_D$ in the proximity, what is the best channel configuration that is to be used by the system such that the UE-1 reaches $\mathcal{Q}$ in the transmissions? We compare the efficiencies among each set of channel parameters through the notion of **Amount of Delay** that is introduced to the system during the transmissions.

We propose to examine the above metrics within the domain defined by the pre-configured SL channel parameters [5] presented in Table I. This is because, in consideration of FR UEs' precious yet limited battery life, it is reasonable to argue that the FR UEs should be set up before being deployed to a disaster site. Even if the settings are to be adjusted on-the-fly, it would still be more power efficient to switch among these pre-configurations. We discuss these pre-configured channel parameters and their impacts in later sections.

To the best of our knowledge, this is the first LTE D2D communications work that focuses on OOC MCC performance in group traffic settings, where the channel parameters are meaningful from 3GPP standard perspective; and could have practical impacts on promoting further research and development of the ProSe in public safety domains.

The rest of the paper is organized as follow. In Section II, we briefly review related work. Then, in Section III and IV, we present our main results on the two identified problems and the methodologies on acquiring these results. Lastly, we conclude our paper and briefly talk about future work in Section V.

II. Related Work

In [2], [3], Griffith et al. analytically modeled the channel reliability problem for PSNs and conducted extensive validations through simulating the SL channels. However, their work did not consider the throughput factor of the UEs; hence has yet to address the impacts of channel reliability onto the higher layers. Also, in their simulations, unlike ours, the channel parameters were rather artificial.

In [6], Cipriano and Panaitopol did investigate the throughput factor; however their work did not focus on the mission-critical aspect. More specifically, both their criteria for a successful transmission and throughput were aggregated over all the UEs. However, we consider the delivery of the leader's message of utmost importance. Also, best-effort throughput is not the most proper metric for limited PSNs, since the extra throughput would be "wasted" after the GBR is met.

To the best of our knowledge, most existing literature discuss performances utilizing the aggregated throughput notion. Hence, by considering the MCC Quality-of-Service perspective, our work is rather novel; and our results will be more prospective and beneficial to the public safety community.

III. Channel Reliability

A. Packet Loss in SL Channels

Now, we briefly review the two factors causing packet losses in SL channels; the general idea is being illustrated in Fig. 2. Then we discuss the steps on how to detect packet losses in the PSCCH and PSSCH.

First of all, since the UEs being considered are using Half-Duplex transmissions, i.e., if a UE is transmitting at one subframe, it won't be able receive at that subframe. Thus, in this case, as shown in the left-most of Fig. 2, the UE with its TB labeled with orange will not be able to hear the leader's transmission at that subframe, and vice versa. However, all the other UEs who are not transmitting at that subframe will be able to hear both.

Secondly, if two UEs' allocated TBs reside in the same subframe and occupy the same set of resources, then a "hard" collision happens to these two set of TBs. In this case, unlike in HDE, not only the two UEs will not be able to hear each other, but also, since these transmitted signals would add up with each other, none of the other UEs who can hear the TBs can differentiate nor successfully decode them. Thus, these two TBs and the encoded bits are to be completely discarded.

There also exists the "soft" collision scenario, where two TBs are partially overlapped. However, since all UEs are within each others' proximity, considering the worst case that the interference would still be too much for anyone to separate the two signals. Hence, these two TBs are still to be discarded. More delicate decisions could be done in the future for this part, such that when the signal-to-interference ratio is high enough, the collided TB(s) could be somewhat recovered.

Also note that, without the instructions encoded in the SCI, UEs will not be able to monitor the correct Physical Resource Block (PRB) spectrum that contains the associated data. Hence, a successful reception of the SCI is the premise of a successful reception of the data.

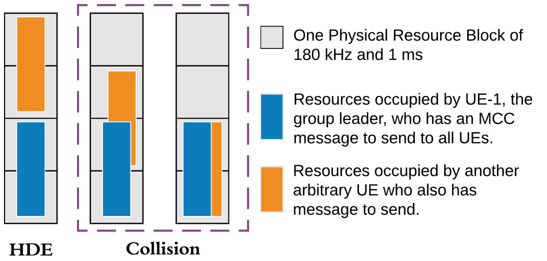

One Physical Resource Block of 180 kHz and 1 ms
Resources occupied by UE-1, the group leader, who has an MCC message to send to all UEs.
Resources occupied by another arbitrary UE who also has message to send.
HDE
Collision

Fig. 2. **Packet Loss Model**: from left to right, each sub-figures showcases the Half-Duplex Effect, the "Soft" Collision, and the "Hard" Collision; where each TB occupies 2 PRBs.

§ Now, for event $\mathcal{S}_C$, in order to detect packet loss in the PSCCH based on our model, we employ the following steps:

1) Since a collision in the PSCCH only happens when two UEs choose the exact same random number n_{PSCCH}, to check for collisions, we simply compare UE-1's random numbers with those of the other UEs. If any UE chooses the same random number as UE-1 does, then, by our assumption, this whole SL period is going to be discarded.

2) Then to check for full HDE, we first compute the SCI allocation subframe numbers (b_1, b_2) from n_{PSCCH} [4] for each UE; and check if there is any UE who has the exact

same subframe pair as UE-1 does; i.e., if any UE will miss both SCI TBs of UE-1.

§ Now, given event $\mathcal{S}_C$ took place in an SL period, for event $\mathcal{S}_D$ in that period, to detect packet loss in the PSSCH, we employ the following steps:

3) First we perform a quick pre-check for full HDE by inspecting if any UE chooses the same I_{TRP} as UE-1 does; i.e., if any UE will miss all four data TBs of UE-1.

4) Then to check for collisions for UE-1, we do the following: for an non-leader UE, at subframes where it shares allocations with UE-1, we check the overlapping situation their TBs in the frequency-domain; if at a subframe these two TBs are colliding, then for UE-1, its TB at this subframe is going to be discarded. We perform this for all non-leader UEs, and check if all of UE-1's TBs are collided.

5) Lastly, for UE-1's remaining non-collided TBs, we iteratively check for HDE by comparing the TRP masks again.

B. PSSCH Reference Measurement Settings

From the packet loss model, it can be seen that given a fixed channel bandwidth and some number of simultaneous transmitters N_u, in the time-domain, the likelihood of HDEs taking places is affected by the lengths of the channels and how the TBs spread out in time within the corresponding channels; while in the frequency-domain, the likelihood of collisions taking places is affected by the allocation size of the TBs.

For SCI messages, 3GPP specifies all the TBs are to occupy only one PRB [4, Table A.6.4-1]; and the spread between the two TBs in time is deterministic. Hence $\mathbb{P}\{\mathcal{S}_C\}$ is only affected by N_u and the length of the PSCCH, L_{PSCCH}.

On the other hand, for data TBs, the *Fixed Reference Measurement Channel for PSSCH* [5, Subclause A.6.5], shown in Table I, defines five different TB schemes with varying Allocated RBs amounts and Modulation of Coding Scheme (MCS) Indices pairs for 10 MHz bandwidth channels. Hence, $\mathbb{P}\{\mathcal{S}_D\}$ is affected by not only N_u, the length of the PSSCH, and $\mathbb{P}\{\mathcal{S}_C\}$, but also the CD (which is an indexing prefix, not an acronym) setting used, as well as k_{TRP}, which stochastically determines how the four data TBs spread in time.

TABLE I
PSSCH REFERENCE MEASUREMENT CHANNEL

Ref. Channel	Allocated RBs	TB Size	MCS Index
$CD.1$	10	872	5
$CD.2$	10	2536	14
$CD.4$	50	12960	14
$CD.5$	2	328	10
$CD.7$	50	25456	23

C. SL Channel Resource Pool Settings

Before the evaluation on channel reliability, we configure our channel RP accordingly. We assume all the 50 PRBs are utilized for SL during public safety events. Also, in the evaluation, we assume all UEs are having good enough signal strengths for demodulating data TBs with varying CD settings; hence we categorize the references into three classes based on the allocation sizes, namely, $CD.5$, $CD.1/2$, and $CD.4/7$. Other RP-related settings are configured as follow:

1. Length of the PSCCH, L_{PSCCH}, is set to 24 subframes per period; based on the result in [2, Fig. 10], such that when $N_u \leq 15$, $\mathbb{P}\{\mathcal{S}_C\}$ is at least the target reliability of 95 %.

2. k_{TRP} is set to either 2 or 4 subframes per TRP mask length (8 subframes). This is because when $k_{TRP} = 1$, too few masks are available, hence the chance of getting HDE is going to be higher; whereas when $k_{TRP} = 8$, HDE will always happen when there are more than 1 transmitters [3].

3. SL Period Length, L_{Period}, is chosen from the values in $\{40, 80, 160, 320\}$ subframes per SL period [7], and is derived from the combined lengths of the PSCCH and PSSCH. Since this factor only affects throughput, for the channel reliability part, we simply simulate under the 40 subframes setting; i.e., there are 16 subframes in the PSSCH, which fits in exactly two sets of TRP masks. This is important since if $k_{TRP} = 2$, two sets of TRP masks are required in order complete one transmission of the four copies of the data TB. Whereas, if $k_{TRP} = 4$, two non-duplicate transmissions can be made.

Later, when throughput is involved, the period length matters in the same way that it determines how many TRP masks there are within one SL period; hence affects the number of non-duplicate transmissions that can be done per period.

D. Evaluation on Channel Reliability

Since currently there is no D2D-enabled chipset publicly available, conducting hardware tests is not an option. Hence, given RP settings, and varying UE and channel parameter settings, we evaluate the channel reliability by conducting Monte Carlo simulations, as follow: $i)$ at each SL period, UEs' SCI TBs and data TBs are generated and allocated in corresponding channels based on the SL UE procedures; and then $ii)$ using the packet loss detection schemes described above, we observe whether events $\mathcal{S}_C$ and $\mathcal{S}_D$ take place at that period. Overall, for each set of channel parameter settings, we check $\mathbb{P}\{\mathcal{S}_D\}$ out of 20 000 (sufficiently large) independent successful $\mathcal{S}_C$ events. In the experiment, all the UEs are transmitting with the same set of parameters. The reliability results over all the possible channel parameter setting combinations are shown in Fig. 3.

In Fig. 3 on the left, we observe that, for a given N_u, the $\mathbb{P}\{\mathcal{S}_C\}$ is not affected by simulation variables. We also note that the PSCCH reliability of at least 95 % with the given RP structure is consistent with the results in [2]. On the right, when $N_u > 2$, different parameter pairs result in drastically varying PSSCH reliability results. As we can observe from the results, by choosing different CD and k_{TRP} pairs, there is a trade-off between having a higher throughput and having a higher reliability. We will address this trade-off in the next section.

Here, we only present the reliability results obtained from one of many runs of our Monte Carlo simulations. The differences among the results of the runs, for the same settings, are no more than 0.5 %. Hence, the results are representative.

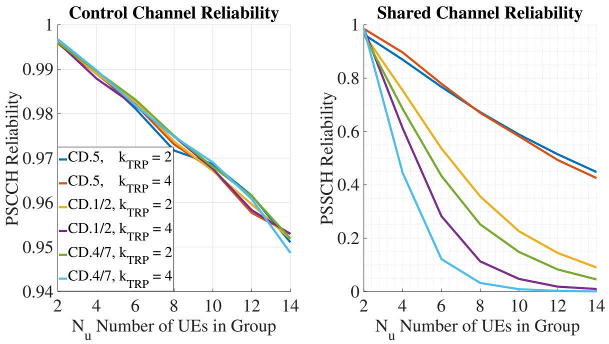


Fig. 3. Simulated results of **PSCCH Reliability** $\mathbb{P}\{\mathcal{S}_C\}$ (left) and **PSSCH Reliability** $\mathbb{P}\{\mathcal{S}_D\}$ (right) for different chosen CD Reference Measurements and k_{TRP} combinations under various N_u values.

IV. Throughput Planning and Delay

In this section, we incorporate the channel reliability results acquired above with the notion of PSN application GBR $\mathcal{Q}$ and try to derive the amount of delay metric. In order to do so, we propose a duplicative transmission scheme that assures the success of group deliveries of each second-worth data; and based on the amount of duplications needed for each SL period under different channel parameter settings, we calculate the amount of delay.

In our experiments, the PSN application GBR values are originally conducted by the Minnesota Department of Public Safety [8] and later modified by the Communications Technology Laboratory (CTL) at the National Institute of Standards and Technology (NIST) with additional network-layer overhead added. Since the reference channel parameters act between the network-layer and the transport-layer, the GBRs we utilize are compatible with the cross-layer settings.

A. Data Broadcasting and Throughput

We observe that, in order for UE-1 to successfully broadcast its application content in the SL channel and attain a throughput at least the amount $\mathcal{Q}$, it needs to transmit in multiple periods within a one second window. The number of transmissions, given an SL RP, then depends on $i)$ the period length L_{Period} and $ii)$ the k_{TRP} setting of the UEs, as explain in the following.

$i)$ For example, if the SL period length is 40 ms, then there are going to be $\lfloor 1000/40 \rfloor = 25$ full SL periods that can be utilized for transmissions; comparing with if the SL period length is 80 ms, then there are fewer, 12, full periods that can be utilized. However, when using a longer period, the data rate (not the actual throughput) is going to increase. When the period length is 80 ms, besides the 24 subframes of control region, the remaining 56 subframes can accommodate 6 TRP masks per period; comparing with only 2 per period in the 40 ms period length case. Later, we denote the number of full periods per second as ρ, and number of TRP masks within one period as γ.

$ii)$ Meanwhile, given the same RP, if $k_{TRP} = 2$, then within each TRP mask, only 1 transmission can be done, since in this case only half of the 4 copies are going to be allocated within each of 8-subframe set; comparing with when $k_{TRP} = 4$, within each period, 2 transmissions can be done, where the second transmission has the exact same allocation as the first one; i.e., the PSSCH reliability stays the same. However, this does not necessarily mean the throughput is increased when using the tighter mask, since as we can see from the reliability results, in most cases, the reliability is significantly lower when $k_{TRP} = 4$ comparing to that when $k_{TRP} = 2$ under the same N_u and CD settings. Thus when $k_{TRP} = 4$, it is more likely to have packet loss taking places, which will jeopardize the whole transmission. Thus, by having a longer period, the penalty for packet loss is going to be higher than that of having a shorter period.

Hence, both factors play important roles in attaining a desired throughput for UE-1, along with the actual allocation choices, i.e., the CD options for PSSCH reference measures.

B. Performance Evaluations

Now we begin to discuss the throughput planning part of our study, where given a fixed number of N_u FR UEs, a PSN application with GBR $\mathcal{Q}$, we want to check if it would be possible for UE-1 to achieve latency-free transmissions, and if not, what is the amount of delay that should be expected?

We introduce the notion of *overall transmission successful probability*, $\mathcal{P}$, as the transmission successful probability over i contiguous periods within each independent second; and if the single period channel reliability is p, then we have $\mathcal{P} = p^i$. We define the target probability as $\hat{p}$; hence as long as $\mathcal{P}$ reaches the target threshold, the second-wise transmission is considered to be successful.

Next, we investigate how the factors mentioned above, i.e., L_{Period}, k_{TRP}, and the selected CD Reference Measurement affect the transmission performance, measured through the delay amount δ. In reality, this evaluation is conducted before the deployment taking place, such that the preferable parameter setups can be implemented for the channel and UEs beforehand; and also after the deployment a UE can switch to a different CD setup that is more appropriate for its own demand based on the obtained results.

We demonstrate the process on how to obtain the amount of delay for an application under various channel and UE settings, by giving a walk-through for the "Next Generation 911 (NG911) Video Medium Resolution" streaming application, with $\mathcal{Q} = 274,562$ bits/sec, when $L_{Period} = 40$ ms, as follow. The formal algorithm description for obtaining the delay amount δ is given by **Algorithm 1**.

1. We first assume all ρ periods within a second are to be used for transmission, to get the minimum throughput that must be met for each transmission and use it to decide what are the choices among the CD settings that are valid for our current parameter settings, by averaging $\mathcal{Q}$ over the total number of transmissions within one second. In the example,

Algorithm 1 Amount of Transmission Delay per Second

```
Input: GBR of Application $Q$, Period Length $L_{Period}$, UEs' $k_{TRP}$,
    Threshold for Overall Transmission Successful Probability $\hat{p}$;
Output: Amount of Delay $\delta$
$\rho \leftarrow \lfloor 1000/L_{Period} \rfloor$;
if ($k_{TRP} == 2$) then
    $\gamma = \lfloor (L_{Period} - L_{PSCCH})/(2 \times N_{TRP}) \rfloor$;
else
    $\gamma = \lfloor (L_{Period} - L_{PSCCH})/N_{TRP} \rfloor$;
end if
$\omega \leftarrow \lceil Q/(\rho \times \gamma \times k_{TRP}/2) \rceil$;
for (each $CD$. Measurement that has TB size $\sigma \geq \omega$) do
    $\eta \leftarrow \lceil Q/(\sigma \times \gamma \times k_{TRP}/2) \rceil$;
    if ($(p)^{\eta} < \hat{p}$) then
        Find smallest $\tau \in \mathbb{N}$, such that $(1-(1-p)^{\tau})^{\eta} \geq \hat{p}$;
        $\delta \leftarrow (\tau \times \eta - \rho) \times L_{Period}$;
    else
        $\delta \leftarrow 0$;
    end if
    if ($\delta < 0$) then
        $\delta \leftarrow 0$;
    end if
    Record current $\delta$ value;
end for
```

there are $\rho = 25$ full periods. If $k_{TRP} = 2$, the number of 16-subframe pairs is $\gamma = \lfloor (40-24)/16 \rfloor = 1$; and if $k_{TRP} = 4$, the number of 8-subframe sets is 2. Then, the total number of transmissions would equal to $(25 \times 1 \times k_{TRP}/2)$.

If $k_{TRP} = 2$, then each period can only allocates one transmission; hence, if all the 25 periods are being used for the transmission, each period must attain a TB of size $\omega = \lceil \mathcal{Q}/25 \rceil = 10983$ bits/allocation, which maps to the $CD.4$ Reference Measurement. Then, if we divide the GBR by the product of the TB size of $CD.4$ and the number of transmissions per period, we would find that if $CD.4$ is used as the allocation scheme, then $\eta = 22$ full periods are actually going to be needed. Here, we assume when all the periods are transmitting, the choice is $CD.4$. If we plan to use $CD.7$ which has a larger TB size, even fewer periods, $\eta = 11$ are going to be needed. Let us denote when $k_{TRP} = 2$ and using $CD.4$ as **Case a)**, and when $k_{TRP} = 2$ and using $CD.7$ as **Case b)**.

If $k_{TRP} = 4$, which doubles the throughput in each period, then the minimum required TB size would be about halved, $\omega = 5492$ bits/allocation, and this again maps to the $CD.4$ Reference Measurement. Similarly, we can choose a higher TB size. If we choose to transmit with $CD.4$, then $\eta = 11$ periods are needed; if $CD.7$ is chosen, then $\eta = 6$. We denote when $k_{TRP} = 4$ and using $CD.4$ as **Case c)**, and when $k_{TRP} = 4$ and using $CD.7$ as **Case d)**.

2. Now, given number of UEs in the group, say, $N_u = 2$, we want to compute $\mathcal{P}$ as p^{η}, i.e., using only η periods.

Hence, for **Case a)**, where $p = 96.41$ %, $p^{11} = 44.74$ %, which is going to be significantly lower than $\hat{p}$, the threshold for having a successful transmission over one second. Similarly, for **Case b)** and **Case c)**, after the exponentiation, we have $0.9641^{11} = 66.89$ % and $0.9861^{11} = 85.75$ %, respectively, and both fall short to meet the goal $\hat{p}$.

However, for **Case d)**, since it requires much less transmission periods, the overall probability, $0.9861^{6} = 91.94$ %, meets the goal, and, as a result would cause no delay at all.

3. For the former three cases, in order to mitigate this exponentially add-up attenuation effect, for each of the periods, we propose that more than one copies are to be transmitted. When the modulated information in a period is transmitted for τ times, call it **trials**, the probability that at least one copy is successfully transmitted is calculated as $(1-(1-p)^{\tau})$.

For **Case a)**, if $\tau = 2$, then the successful probability over a second is $(1-(1-p)^{\tau})^{11} = 97.20$ %, and thus the success of transmissions is going to be assured. This is very similar to the idea of transmitting multiple copies within a period, but now the duplication is regarding the periods instead of the allocations within the periods. For the other cases, $\tau = 2$ as well.

4. However, as we just discussed, by transmitting extra copies of the periods, the delivery for the content within each second is going to be delayed.

In both **Case b)** and **Case c)**, since all the 11 periods are going to be used for transmissions, and each periods are going through 2 **trials** to assure the success of transmission, in total $11 \times 2 = 22$ periods are going to be transmitted, which still falls into the one second (25 periods) range; and hence also would not cause any delay. On the other hand, for **Case a)**, 44 periods are needed, where 25 of them fall within the one second range, and the other 19 are going to be counted as the delay portion; hence, for the content of the first second, the delay is going to be $\delta = 19 \times 40 = 760$ ms.

Hence, delay-wise, we can see **Case a)** would be the least favorable scheme among the four to choose for this application, under currently tested settings. For **Case d)**, it can also transmit multiple copies to further improve its $\mathcal{P}$.

In general, by adjusting the three parameters, which results in different η and τ values, will change the amount of delay. Hence, for various PSN applications with different GBR values, for various N_u values, and among all the valid combinations of the three parameters, we conduct extensive experiments to collect the resulting delay amount δ using our proposed **Algorithm 1**. Our experiments generate sets of tables on the δ values with completeness; hence network operator can then set up the SL channel and the D2D UEs according to these pre-computed δ values so that the delay budget is met. The results can also be used to draft new latency budget policies, since none of the existing ones were made for OOC scenario in particular.

Due to space limitation, we only present part of our experimental results, as shown in Fig. 4 for the "NG911 Video MQ" application which has a relatively high GBR value; and in Fig. 5 for the "Phone Voice" application who has the intrinsic highest priority in communication, with a moderate GBR of 30,440 bits/sec. and for both applications we focus on $N_u = 4$, where zero delay is achievable. When N_u becomes larger, it is most likely that at least some degrees of delay is to be introduced, and network operators can evaluate which schemes are better under circumstances by considering the trade-offs.

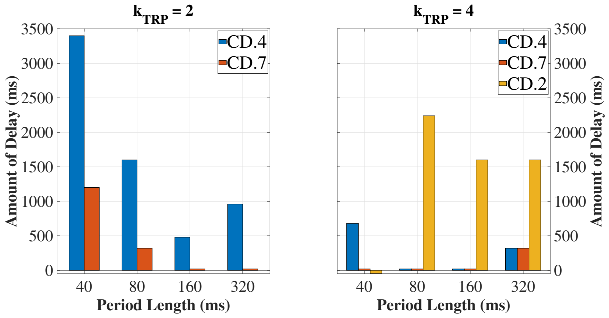


Fig. 4. Amount of delay for “NG911 Video MQ” application with $\mathcal{Q} = 274,562$ bits/second, for $N_u = 4$; within each group of bars, from left to right, each bar represents $CD.4$, $CD.7$ and $CD.2$ (if exists), respectively.

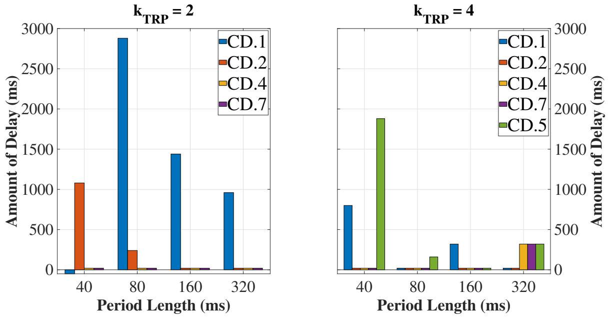


Fig. 5. Amount of delay for “Phone Voice” application with $\mathcal{Q} = 30,440$ bits/second, for $N_u = 4$; within each group of bars, from left to right, each bar represents $CD.1, 2, 4, 7$, and $CD.5$ (if exists), respectively.

In both Fig. 4 and Fig. 5, we use bars with very small heights to represent that there are no delays for the communications under certain parameter settings. Also, the yellow and blue bars with negative values indicate invalid data, since those settings will not be able to achieve $\mathcal{Q}$ at $L_{Period} = 40$ ms.

When comparing the two figures, we can see that when $\mathcal{Q}$ becomes smaller, more reference measurements with smaller TB sizes become valid for transmission. However, it is obvious that, in most cases, although channel parameters with smaller TB sizes have much higher channel reliability, they do **NOT** produce smaller delay values, due to the exponentially add-up attenuation on the reliability. Two exceptions being that, in Fig. 5 when $L_{Period} = 160$ and 320, settings with smaller TB sizes would actually produce smaller amount of delays. On the other hand, channel parameters with larger TB sizes usually require receiving UEs to have better channel qualities in order to actually demodulate the data TBs, which obviously is not always achievable in real-life situations. Thus, it is up to network operators to balance the trade-off accordingly.

V. Conclusion and Future Work

In this work, we thoroughly investigate the channel parameters that will affect the channel reliability, and consequently, the amount of delay introduced during communication, with varying FR group sizes and GBRs of PSN applications. We introduce an approach to assure the successful delivery of content under a rigorous packet loss model, and derive the amount of delay introduced by applying this approach under all the possible parameter combinations.

It can be seen from the amount of delay results that, although in general, settings that carry out larger TB sizes per allocation, despite of having much lower channel reliability, would produce smaller amount of delays; exceptions do exist, and hence the problem is indeed worth studying. Thus, our results have the potential to benefit network operators and researchers in the future operations and studies on SL channels.

As we all know, standards are evolving as technologies develop. Therefore, some of the parameters we investigate in this work may be extended and end up having more values to be utilized to increase the channel reliability, UE throughput, and thus reduce the amount of delay being introduced. We plan to generalize the domains of certain parameters, which will likely make it impossible to construct a table that contains all the parameter information. We will investigate the development of an algorithm that would yield an optimal setting without having to generate the complete table.

Furthermore, there are plentiful of topics that can be expanded into and studied based on our work. For example: **Role-based D2D UEs** where UEs may choose different parameters based on their demand and behavior/roles; and, **Multi-hop D2D Communication** where messages being passed through link paths between base station and remote UEs via multiple D2D-enabled relay UEs.

Acknowledgment

This work was in part supported by the National Institute of Standards and Technology (NIST) through award 70NANB16H021.

References

[1] Federal Communications Commission (FCC). (2017) Communications Status Report For Areas Impacted By Tropical Storm Harvey. https://www.fcc.gov/harvey. [Online; accessed 2019].

[2] D. W. Griffith, F. J. Cintrón, and R. A. Rouil, “Physical sidelink control channel (PSCCH) in mode 2: Performance analysis,” in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–7.

[3] D. Griffith, F. Cintron, A. Galazka, T. Hall, and R. Rouil, “Modeling and simulation analysis of the physical sidelink shared channel (PSSCH),” in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–7.

[4] 3GPP, “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.213, July 2018, version 14.7.0.

[5] ——, “Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.101, July 2018, version 14.8.0.

[6] A. M. Cipriano and D. Panaitopol, “Performance analysis of sidelink data communications in autonomous mode,” in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018, pp. 1–6.

[7] 3GPP, “Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification,” 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.331, January 2019, version 14.9.0.

[8] Minnesota Department of Public Safety, “Public Safety Wireless Data Network Requirements Project Needs Assessment Report Phase 1-Task 4/Deliverable 2,” Tech. Rep., May 2011.

Feng, Siyuan; Choi, Hyong-Ah; Griffith, David W.; Rouil, Richard A. ”On Selecting Channel Parameters for Public Safety Network Applications in LTE Direct.” Paper presented at IEEE Consumer Communications & Networking Conference (CCNC), Las Vegas, NV. January 10, 2020 - January 13, 2020.

Throughput Analysis between Unicast and MBSFN from Link Level to System Level

Chunmei Liu*, Chen Shen†, Jack Chuang*, Richard A. Rouil*, and Hyeong-Ah Choi†
*Wireless Networks Division, National Institute of Standards and Technology, USA
†Department of Computer Science, George Washington University, USA
Email: *{chunmei.liu, jack.chaung, richard.rouil}@nist.gov, †{shenchen, hchoi}@gwu.edu

***Abstract*—Public safety incidents typically involve significant amount of group traffic. This paper initializes our study in exploring the potential spectrum savings and improvement in first responders experience by using Multicast Broadcast Single Frequency Network (MBSFN) to serve group traffic among first responders. Towards this goal, this paper proposes a comprehensive methodology that closely follows The 3rd Generation Partnership Project (3GPP) specifications and considers unicast multiple-input multiple-output (MIMO) and MBSFN without MIMO. High fidelity Block Error Rate (BLER) curves for both MBSFN and unicast are generated, and Signal-to-noise ratio (SNR) points to switch Channel Quality Indicators (CQIs) are extracted and analyzed. Several simulation scenarios have been designed and the empirical results are analyzed.**

***Index Terms*—BLER curves, CQI, LTE, MBSFN, MIMO, public safety broadband network, unicast.**

I. Introduction

In 2012, the Congress passed the Middle Class Tax Relief and Job Creation Act of 2012 [1]. The act calls for the establishment of a national public safety broadband network to expand high-speed wireless broadband and to improve communications interoperability among first responders. Meanwhile, the Long Term Evolution (LTE) networks have been deployed for several years throughout the nation to serve millions of commercial users, and point-to-point unicast transmission is the typical mechanism deployed.

Compared with commercial traffic, public safety incidents typically involve significant amount of group traffic among first responders [2]. Using traditional point-to-point unicast communication to serve this traffic requires to transmit the same content multiple times over the air interface. On the contrary, if multicast transmission is used, the content needs to be transmitted only once over the air interface. This potential saving on the precious spectrum and improvement in first responders' user experience trigger the exploration of using multicast to serve group traffic among first responders. Towards this goal, this paper considers the public safety broadband network that is currently built upon LTE technologies using Band 14 (B14) 10 MHz bandwidth Frequency Division Duplex (FDD), with focus on downlink and Multimedia Broadcast Multicast Service Single Frequency Network (MBSFN) [3].

MBSFN has been studied in the literature from various perspectives. In [4], four adaptive Modulation and Coding Scheme (MCS) selection algorithms have been proposed. Except the bottom-up algorithm which selects the minimum MCS for all user equipments (UEs) in MBSFN, the others cannot guarantee all UEs to successfully decode the data. Also this work directly uses unicast Block Error Rate (BLER) curves for MBSFN. In [5], MBSFN and Single Cell Point To Multipoint (SC-PTM) are evaluated together with the unicast transmission. However, an important feature, multiple-input multiple-output (MIMO), is not considered in unicast. In [6], an optimal UE grouping algorithms in MBSFN is proposed. Since the minimum MCS is selected for all UEs in MBSFN, UE grouping with separate assigned resource may improve the overall performance. There the UE is associated with the same MCS in unicast and multicast modes. As we will see later, it is not always the case. The performance of Multimedia Broadcast Multicast Services (MBMS) and unicast when transmitting video is compared in [7]. The authors use average MCS of all UEs for the expected performance calculation. The comparison is also based on the same MCSs for unicast and multicast.

Our ultimate goal is to quantify network and user performance improvements if MBSFN and unicast communications are adapted based on specific network deployments and device distributions. Our system model and simulation closely follow the 3rd Generation Partnership Project (3GPP) standard with MIMO features included for unicast transmission but not MBSFN [3] [8]. Signal-to-interference-plus-noise ratio (SINR) values in unicast and MBSFN are calculated separately based on their unique models, thus we do not simply use MCS in unicast settings for the MBSFN. In addition, by considering comprehensive physical (PHY) and medium access control (MAC) layers, link level simulations were conducted and traceable high fidelity BLER vs Signal-to-Noise Ratio (SNR) vs Channel Quality Indicator (CQI) curves were generated and analyzed. To our knowledge, this modeling with the above features together for system level performance analysis is the first of its kind.

The rest of this paper is organized as follows. Section II reviews MBSFN and unicast as defined in 3GPP. Section III outlines our methodology and describes our link level simulation design in detail. Simulation results and analysis are presented and summarized in Section IV. Finally, Section V summarizes the paper and outlines steps for future research.

II. Review of unicast and MBSFN in 3GPP

With focus on major factors that have direct impacts on performance comparison and simulation design, this section

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Throughput Analysis between Unicast and MBSFN from Link Level to System Level." Paper presented at 2019 IEEE 90th Vehicular Technology Conference: VTC2019-Fall, Honolulu, HI, US. September 23, 2019 - September 25, 2019.

reviews MBSFN as specified by 3GPP, and highlights differences between unicast and MBSFN.

In MBSFN, MBMS data is transmitted from multiple cells to the destination UE. All cells involved are tightly synchronized and transmit the same content over the same subcarriers, and all transmissions received by the UE differ only in arrival times, amplitudes, and phases. Given the redundant transmissions via multiple eNodeBs, the SINR is improved, with the greatest improvement being seen at the cell edge.

Fig. 1 illustrates a typical MBSFN subframe as specified in 3GPP [9], with one subframe lasting one Transmission Time Interval (TTI) in time domain. The non-MBSFN region is for controls and signals, and the MBSFN region is where MBMS data is transmitted. In the MBSFN region, in addition to MBMS data, MBSFN reference signals (RSs) are used by the UEs for channel estimation. Note that there is a gap between the two regions [9].

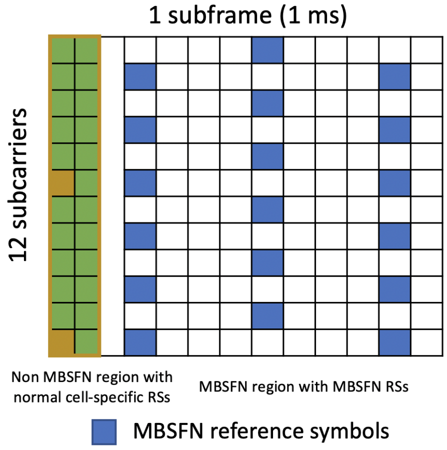

1 subframe (1 ms)
12 subcarriers
Non MBSFN region with normal cell-specific RSs
MBSFN region with MBSFN RSs
MBSFN reference symbols

Fig. 1: MBSFN Subframe Structure

As a comparison, Fig. 2 illustrates a typical unicast subframe. The figure shows the configuration with normal cyclic prefix (CP) and 2 symbols for control region, which are typical LTE network settings. The cell-specific reference signals (CRSs) shown are with one antenna port. Subframe structure with other configurations can be found in [9].

By comparing Fig. 1 and Fig. 2, it can be seen that MBSFN RSs have tighter spacing than unicast CRSs, which leads to higher overhead and less resource elements (REs) available for user data transmissions. In addition, MBSFN subframes and non-MBSFN subframes are interleaved in time. Moreover, while normal CP is used in typical LTE networks for unicast, extended CP is used for MBSFN to facilitate transmissions from multiple cells to arrive at the UE within CP. This reduces the number of OFDM symbols available for user data from 12 for unicast to 10 for MBSFN, as shown in Fig. 1 and Fig. 2. Furthermore, since subframe 0, 4, 5, and 9 in each radio frame carries signals that are essential for network operations in their data region [9], these four subframes are reserved for unicast transmission and cannot be configured as MBSFN subframes.

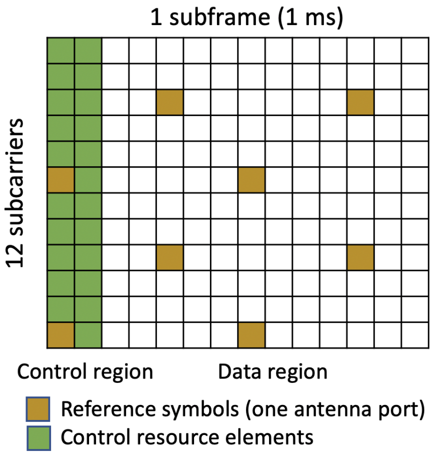

1 subframe (1 ms)
12 subcarriers
Control region
Data region
Reference symbols (one antenna port)
Control resource elements

Fig. 2: Unicast Subframe Structure - One Antenna Port

3GPP also specifies that a single transport block (TB) is used per subframe for MBSFN, and that TB uses all the MBSFN resources in that subframe [3]. Specific to the public safety broadband network on B14 with 10 MHz spectrum, this corresponds to 50 resource blocks (RBs). Whereas in unicast, RBs in one subframe are typically divided into multiple sets, with each set assigned to a different user.

For unicast, MIMO has been one major technology adopted by LTE and provides significant improvements in performance and spectral efficiency. Up to release 15, 3GPP defines 10 transmission modes that are intended to fully utilize MIMO under different situations [8]. Up to two codewords can be used for a single UE. Contrarily, in MBSFN, due to its multicast nature, 3GPP specifies no transmit diversity scheme, and MBSFN is mapped on a single layer spatial multiplexing. That is, MBSFN cannot take advantage of MIMO gains.

In addition, hybrid automatic repeat request (HARQ) is a key component that ensures reliability and performance for unicast. In MBSFN, 3GPP specifies that a single transmission is used, and there is no Radio link control (RLC) retransmissions and no HARQ. Hence in order to deliver acceptable service to upper layers, lower target BLER for single transmission is typically used for MBSFN. Lost MBMS packets for certain UEs are typically recovered by retransmissions from the application layer via unicast.

III. Methodology

This section outlines the overall design for performance comparison between unicast and MBSFN, then describes the link level simulation design in detail.

A. System Model

Fig. 3 illustrates our system design. After setting up the scenario to be simulated, both unicast and MBSFN are run, and the resulting performance is recorded and compared. For both unicast and MBSFN, without loss of generality, perfect channel knowledge and zero feedback delay are assumed.

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Throughput Analysis between Unicast and MBSFN from Link Level to System Level." Paper presented at 2019 IEEE 90th Vehicular Technology Conference: VTC2019-Fall, Honolulu, HI, US. September 23, 2019 - September 25, 2019.

The control signaling is not modelled. For MBSFN, the unicast traffic used for application layer retransmissions are not modelled.

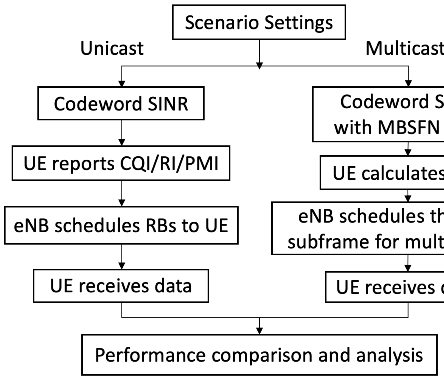


Fig. 3: Overall Design

For both unicast and MBSFN, due to its reputation of accurately predicting BLER [10]–[12], Mutual Information based Exponential SNR Mapping (MIESM) is applied to map the post-equalization SINR to Additive white Gaussian noise (AWGN) equivalent effective SNR. The general equation of the SINR to SNR mapping is as below:

$$\beta_m = f_m^{-1}\Big(\frac{1}{N_{RE}}\sum_{c=1}^{N_{RE}} f_m(\beta^c)\Big) \tag{1}$$

where β_m is the AWGN equivalent SNR for modulation m, $f_m(.)$ is the Bit Interleaved Coded Modulation (BICM) capacity of modulation alphabet m that is associated with the MCS m, β^c is the post equalization SINR for RE c, and N_{RE} is the total number of REs utilized for the transmission. In our simulation, to save simulation time, the averaging is from half RB to codeword [8].

By using Eq. (1), we can translate any channel-interference scenario to an equivalent AWGN channel, and the sample based simulation is required only once when acquiring BLER vs SNR vs CQI curves, which will be described in detail in Section III-B.

For unicast, the overall channel consists of the precomputed pathloss, shadowing, and microscale fading. The post-equalization SINR is calculated using zero forcing equalizer. Then the MIESM maps the time-frequency selective channel experienced over a number of REs and over spatial streams when spatial multiplexing is employed, to an AWGN channel that achieves the same average spectral efficiency in terms of the BICM capacity. Next, based on the AWGN BLER curves from link level simulation, each UE estimates the achievable throughput for all possible number of layers, and selects the optimal Rank Indicator (RI). Afterwards, each UE reports the selected CQI, RI, and Precoding Matrix Indicator (PMI) to the eNodeB. With the above reports from the UEs, the eNodeB schedules its RBs based on the configured scheduler, where proportional fairness scheduler is selected for our study. Finally, with the BLER curves, post-equalization SINR, and the assigned MCS from the eNodeB, each UE records successful packets and sends NACK for unsuccessful packets.

For MBSFN, the channel model used is the same as the unicast one. While calculating the post-equalization SINR, we use run time precoding and Zero-Forcing equalizer, with the modelled constructive signals from eNodeBs in the MBSFN area. The detailed modeling will be explained in [13]. The resulting post equalization SINR for RE c, β^c, can be calculated as in Eq. (2) and (3), where I_M represents the inter-symbol interference from cells that participate MBSFN transmission.

$$\beta^c = \frac{1}{\sum_{i=1}^{N_M}(1-w_i)P_i^c\|F^cH^{ic}\underline{1}_{nTx}\|^2 + I_M + \|F^c\|^2 N_0} \tag{2}$$

$$I_M = \sum_{i=N_M+1}^{N} P_i^c\|F^cH^{ic}W^{ic}\|^2 \tag{3}$$

where N is the total number of cells, and the first N_M cells are those that participate in MBSFN transmission, and cell $N_{M+1}, ..., N$ are cells that do not participate in MBSFN transmission. w_i is the weighting function for cell i as defined in [14]. $P_i^c, i = 1, ..., N$ is the signal power from cell i at RE c after taking into account path loss and shadowing but without microscale fading. $H^{ic}, i = 1, ...N$ is the channel matrix from cell i to the UE at RE c. W^{ic} is the precoding matrix used by cell $i, i = N_M + 1, ..., N$. F^c is the zero-forcing equalizer. N_0 is the thermal noise. $\underline{1}_{nTx}$ is (nTx)x1 column vector with all elements being 1.

The MIESM approach is then applied to map the post equalization SINR to AWGN equivalent SNR, and each UE calculates it CQI with the MBSFN BLER curve from link level simulation (Section III-B) and reports it back to the eNodeB. The MBSFN eNodeBs select the MCS based on the minimum CQI among all MBSFN UEs and assign the whole subframe for multicast data. At last, UEs receive the data and calculate their throughput.

B. Link Level Simulation Design

As mentioned in Section III-A, the objectives of the link level simulation are to generate BLER curves for each CQI under AWGN channel, for both unicast and MBSFN, and to retrieve CQI switching points associated with target BLERs. For this purpose, Vienna link level simulator [10][1] is selected as the base platform. The simulator is designed for unicast. Its PHY layer architecture is consistent with the typical LTE physical layer procedure and that defined in 3GPP [9]. After evaluation, our simulation reused the simulator for unicast with little modifications on settings. For MBSFN, the simulator is extended to support MBSFN.

Specifically, the link simulated for both unicast and multicast is one point-to-point link between one eNB and one

[1] Any mention of commercial products in the paper is for information only; it does not imply recommendation or endorsement by NIST.

UE, with public safety band downlink center frequency of 763 MHz and 10 MHz bandwidth. CQI indices are the ones based on QPSK, 16QAM, and 64QAM, as shown in Table 7.2.3-1 in [8]. AWGN channel is chosen as mentioned in Section III-A. At the UE side, perfect channel knowledge is assumed and zero-forcing equalizer is employed. At the eNB side, transport block size (TBS) is calculated based on MCS order, coding rate, and RBs allocated as in Table 7.2.3-1 in [8]. In addition, in case of large TBS, TB segmentation and de-segmentation are performed to fit the maximum code block size defined per 3GPP [15]. Moreover, since the BLER curve is to be used in later system level simulation (Section III-A), there is no HARQ retransmission in link level simulation.

For unicast, normal CP is selected, and number of OFDM symbols for control region is set to be 2, both of which are typical LTE network deployment settings (Fig 2). Since the focus is on user data transmissions, control region is not simulated. However, since primary and secondary synchronization signal (PSS and SSS) fall into data region and occupy resource elements (REs), they are both simulated (Section II). Regarding CRSs, different RS sets map to different antenna ports [9]. To reduce the number of BLER curves used in later system level simulations and without loss of generality, RSs mapping to one antenna port is selected for this study (Fig. 2).

Per 3GPP standards, for unicast, when UE reports CQI, transport block error probability should not exceed 0.1 [8]. Hence the target BLER for CQI switching is selected to be 0.1. Also note that since PSS and SSS fall into data region in subframe 0 and 5 but not in other subframes [8], the 10 subframes within one radio frame are no longer identical in the sample space. However, within one subframe, the number of REs occupied by primary and secondary synchronization signals is small compared with total number of REs, and their impacts on BLER is hence insignificant. Therefore, to simplify the design, all subframes are treated equally in BLER calculation. 5000 subframes are simulated to provide enough confidence level.

For MBSFN, extended CP is selected as specified by 3GPP [9]. Similar to unicast simulation design, since the focus is on user data transmissions, non-MBSFN region is not simulated (Fig. 1). As mentioned in Section II, subframe 0, 4, 5 and 9 cannot be used for MBSFN. Hence they do not contribute to BLER for MBSFN and therefore are not simulated. Different from unicast, there is no need to simulate PSS and SSS. This is because they fall into subframe #0 and #5, both of which can not be used as an MBSFN subframe.

Since in MBSFN there are no HARQ retransmissions and no RLC retransmissions, lower target BLER is expected and selected to be 0.01. In addition, MBSFN subframes are identical in sample space, and 5000 subframes are simulated to provide enough confidence level.

One challenge in MBSFN simulation design is the location of MBSFN RSs. As shown in Fig. 1, MBSFN RS distribution is not the same across two slots. However, Vienna link level simulator is structured in the way that a transmit signal is mapped on REs based on symmetric slots. One major function of RSs is to support channel estimation. Since our simulation converts SINR in multipath channel to equivalent SNR and perfect channel knowledge is assumed, the locations of the RSs are not expected to change the BLER results. Based on this analysis, one design to simulate the MBSFN RSs is to simulate them by their total number instead of actual locations in subframes. We tested this design by comparing BLER curves of 16 RSs, one with uniformly distributed 16 RSs across the subframe, and the other with concentrating the 16 RSs in two OFDM symbols. The resulting BLER curves show no observable significant differences.

Based on the former analysis and verification, MBSFN RSs are simulated with focus on their total number and by putting 9 RSs at subcarrier 2 to 10 for symbol 3, and another 9 RSs at subcarrier 2 to 10 for symbol 9. Moreover, because of the assumption of perfect channel knowledge, the sequences used by MBSFN RSs would not change the simulation results. Hence, to save the effort in generating MBSFN RS sequence, the reference signal sequences Vienna generated for unicast are reused.

IV. Simulation results

Based on the simulation design described in Section III, link level simulations have been implemented and performed for the public safety band. The obtained BLER curves are shown in Fig. 4 and Fig. 5 for unicast and MBSFN, respectively. The curves in Fig. 4 and Fig. 5 successfully support our initial projection that there would be no significant difference between MBSFN and unicast BLER curves over SNR in AWGN channel, despite that the MBSFN BLER curves shift a little bit to the left. The CQI is the major parameter which will impact BLER in AWGN.

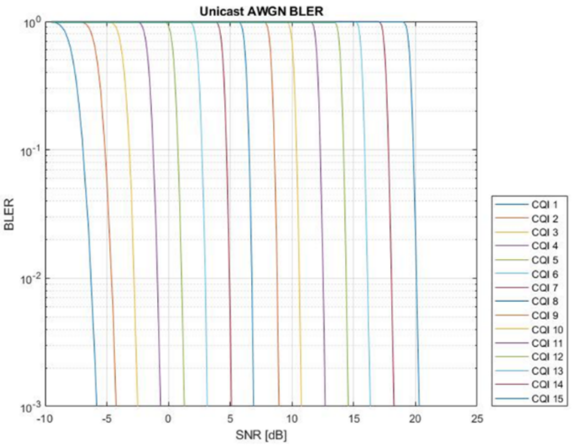


Fig. 4: Unicast BLER for AWGN B14 700 MHz Downlink

Given a target BLER of 0.1 for unicast and 0.01 for MBSFN, we also extracted CQI switching points and listed them in Table I. For comparison purposes, switching points for target BLER of 0.01 for unicast and 0.1 for MBSFN are also listed. We discovered that for the same target BLER, for

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Throughput Analysis between Unicast and MBSFN from Link Level to System Level." Paper presented at 2019 IEEE 90th Vehicular Technology Conference: VTC2019-Fall, Honolulu, HI, US. September 23, 2019 - September 25, 2019.

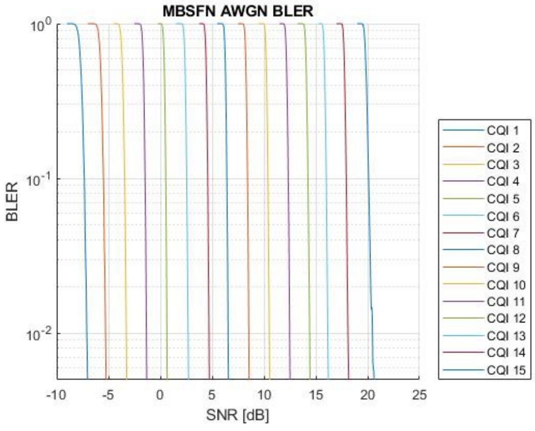


Fig. 5: MBSFN BLER for AWGN B14 700 MHz Downlink

low CQIs, MBSFN requires less SNR than unicast, while for high CQIs, unicast requires less SNR. We also discovered that with different target values, the switching points do not differ significantly.

Note that similar BLER curves and CQI switching points do not imply that the unicast and MBSFN will perform similarly in actual networks. Due to different system architecture, in actual networks when unicast or MBSFN is applied, the first responders could experience different combined SINR before the equalizer even in the same multipath scenario (for example after MBSFN gain especially at cell edge). After the SINR to SNR translation, the AWGN equivalent SNR is different. This could lead to different CQI reports for MBSFN and unicast to meet BLER requirement, and hence different performance first responders would experience. This projection is verified by the system level simulation results described below.

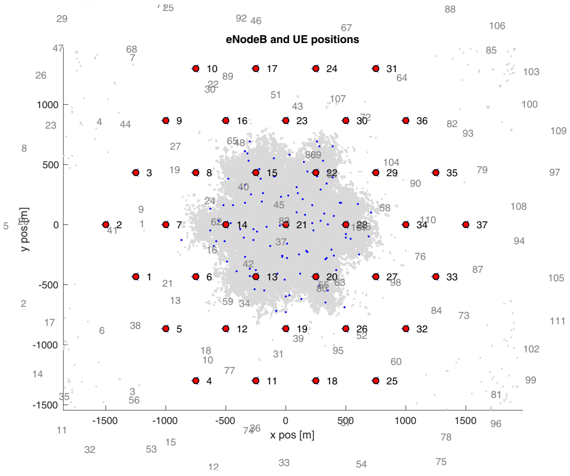


Fig. 6: Network Layout

The network layout used in the system level simulation is shown in Fig. 6. The network is the typical tri-sectored site hexagonal grid with 37 sites and 111 cells in total. First responders are uniformly located in the shaded center 7 sites 21 cells, with 5 first responders per cell unless mentioned otherwise. The outer two rings of 30 sites are included to simulate interference. Inter-site distance is set to be 500 meters. The eNB transmission power is set as 40 W, and the antenna height and the receiver height are 32 m and 1.5 m, respectively. The number of eNB transmit antennas is 8, and the number of UE receiver antennas is 4. Urban macro channel model is used [16]. For unicast, transmission mode 9 [8] is employed, which allows spatial multiplexing to take advantage of MIMO gains. For MBSFN, the MBSFN area is the center 7 sites where first responders are located.

Fig. 7 shows the cumulative distribution function (CDF) of the average AWGN equivalent effective SNR over time experienced by codewords for unicast and MBSFN. Note that in unicast when more than 1 layer is applied (MIMO), the number of codewords could be 2. In this case, the effective SNR of each codeword contributes to one sample in generating the CDF. This figure shows that MBSFN has significant higher effective SNR than unicast. This improvement comes from several factors. The first is that in unicast the cells in the center 7 sites generate interference to each other, while in MBSFN they do not. The second is the MBSFN gain due to transmissions from multiple cells. The third is that in unicast when there are two codewords, the signal power is distributed to two codewords instead of one, which lower the SNR of each codeword.

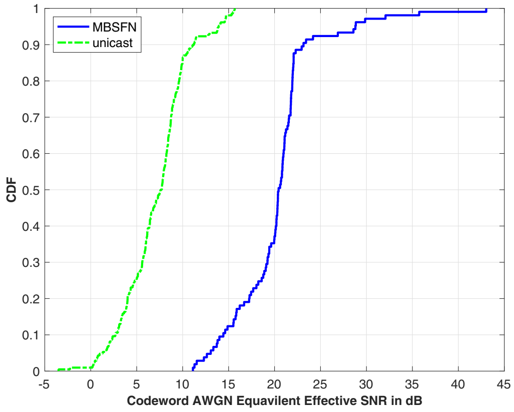


Fig. 7: CDF of AWGN Equivalent Effective SNR

Fig. 8 shows comparison of reported CQI between MBSFN and unicast from 10 randomly selected UEs in TTI 23. Note that for unicast, some UEs report 2 CQIs and some report 1. This is because with MIMO when there are two codewords, each codeword is associated with one CQI. It can be observed that in the selected case, MBSFN has higher CQI than unicast, which is consistent with Fig. 7. Contrary to the assumption in some literature which assume that MBSFN MCS is the

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Throughput Analysis between Unicast and MBSFN from Link Level to System Level." Paper presented at 2019 IEEE 90th Vehicular Technology Conference: VTC2019-Fall, Honolulu, HI, US. September 23, 2019 - September 25, 2019.

	CQI	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MBSFN	target BLER 0.1	-7.3	-5.45	-3.4	-1.45	0.55	2.55	4.6	6.45	8.45	10.4	12.4	14.3	16.1	18	20.1
	target BLER 0.01	-7.05	-5.25	-3.25	-1.3	0.65	2.7	4.7	6.55	8.55	10.6	12.5	14.5	16.2	18.2	20.5
Unicast	target BLER 0.1	-6.94	-5.15	-3.18	-1.25	0.76	2.7	4.69	6.53	8.57	10.4	12.3	14.2	15.9	17.8	19.8
	target BLER 0.01	-6.29	-4.62	-2.8	-0.91	1.06	2.98	4.93	6.75	8.79	10.6	12.5	14.4	16.1	18.1	20.1

TABLE I: CQI switching points at target BLER 0.1 and 0.01, SNR in dB

minimum of unicast MCSs of all UEs, the lowest MBSFN CQI is not necessarily the lowest unicast CQI.

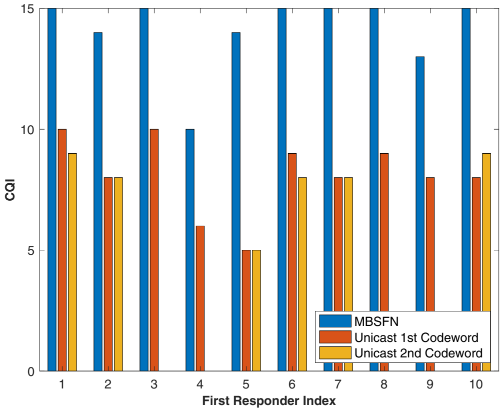


Fig. 8: Comparison between MBSFN CQI and Unicast CQI

The CDF of first responder throughput is shown in Fig. 9, for both unicast and MBSFN. Note that different from Fig. 7, in unicast when spatial multiplexing is triggered and 2 codewords are used for the UE, both codewords contribute to this UE's throughput and together generate one sample in throughput CDF. Meanwhile, in MBSFN 6 out of 10 subframes are used, which is the maximum specified by the standard. It can be observed that the MBSFN CDF is close to a vertical line. This is because of the low target BLER of 0.01, and the fact that all UEs are receiving the same content at the same time, using the same MCS and RBs. Hence all first responders experience similar throughput. In this specific case, MBSFN shows higher throughput. MBSFN also has higher average throughput of 9.76 Mb/s than unicast 6.02 Mb/s, which is not shown in the figure. With increasing number of first responders, this throughput gap will be bigger.

To show the impact of selecting the lowest MCS in MBSFN, the potential throughput UEs would achieve per their individual CQI reporting is plotted in red. This potential throughput is much higher than the actual throughput, and its lowest value is the same as the actual MBSFN throughput. The potential MBSFN throughput is also much higher than the unicast throughput. One reason is the higher effective SNR as shown in Fig. 7. Another reason is that MBSFN uses all RBs for its transmission and that applies to all UEs, while in unicast the RBs are shared by the UEs. That is, the resource used by each UE is much more in MBSFN than that in unicast. However, given the lowest MCS selection in MBSFN, the actual throughput experienced by first responders is much lower than the potential MBSFN throughput.

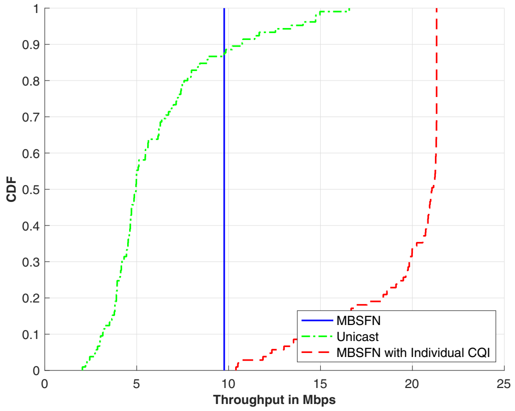


Fig. 9: Throughput CDF for Unicast and MBSFN

Fig. 10 shows the histogram of CQIs for MBSFN reported from all UEs in one TTI. In this case, although majority of the UEs reported the highest possible CQI of 15, there are UEs that reported CQI of 10. These UEs lowered the MCS used by MBSFN, and hence the MBSFN throughput.

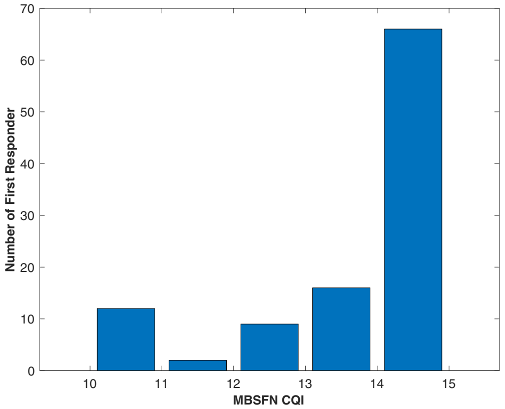


Fig. 10: MBSFN CQI Histogram in TTI 23

Fig. 11 further shows how minimum CQI in MBSFN varies in time. It can be observed that in most time, the minimum

MBSFN CQIs remain stable. The more first responders per sector, the lower the minimum CQI is.

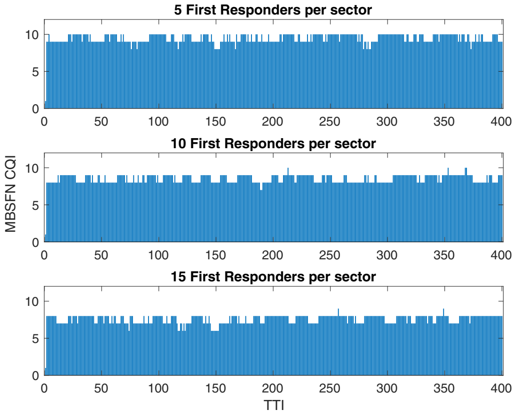


Fig. 11: MBSFN Minimum CQI with Time

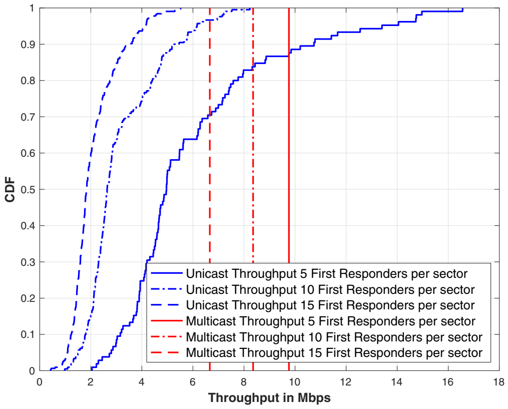


Fig. 12: Throughput CDF with 5, 10, and 15 First Responders per Sector

Finally, Fig. 12 shows first responders' throughput CDF when there are 5, 10, and 15 first responders per sector, respectively. It can be observed that the first responders' throughput in unicast decreases significantly as first responders' density increases, while in MBSFN it is less sensitive. The relatively lower throughput in MBSFN with more first responders is due to the slightly lower MBSFN CQI, as shown in Fig. 11. Also note that in the cases shown, MBSFN throughput is higher than majority of unicast throughput. In case of 15 first responders per sector, the MBSFN throughput is even higher than the highest throughput in unicast. This is consistent with our analysis in Fig. 7.

V. Conclusion

In this paper we generated high fidelity BLER curves for both MBSFN and unicast, and extracted and analyzed SNR points to switch CQIs. The BLER curves and CQI switching points were then fed into our comprehensive system level simulations, which considers unicast with MIMO and MBSFN without MIMO as well as all 3GPP PHY and MAC features. The resulting SINR, throughput, and CQI for both unicast and multicast were also analyzed.

While throughput is one major metric for performance, another perspective of group traffic is that the content is available to first responders simultaneously. Our next step is to introduce other metrics in addition to throughput, and to use the results in this paper to further investigate the trade offs between unicast and multicast. We will also optimize LTE system to meet first responder needs and to ensure quality service for everyone.

References

[1] FirstNet. https://firstnet.gov/content/firstnet-building-nationwide-public-safety-network.

[2] "Minnesota Department of Public Safety Public Safety Wireless Data Network Requirements Project Needs Assessment Report," Minnesota PSN, Tech. Rep., May, 2011.

[3] 3GPP TS36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," 3GPP, Standard.

[4] A. Alexiou, C. Bouras, V. Kokkinos, A. Papazois, and G. Tsichritzis, "Modulation and Coding Scheme Selection in Multimedia Broadcast over a Single Frequency Network Enabled Long Term Evolution Networks," *International Journal of Communication Systems*, vol. 25, 12 2012.

[5] A. Daher, M. Coupechoux, P. Godlewski, P. Ngouat, and P. Minot, "SC-PTM or MBSFN for Mission Critical Communications?" in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, June 2017, pp. 1–6.

[6] J. Chen, M. Chiang, J. Erman, G. Li, K. K. Ramakrishnan, and R. K. Sinha, "Fair and Optimal Resource Allocation for LTE Multicast (eMBMS): Group Partitioning and Dynamics," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 1266–1274.

[7] S. Mitrofanov, A. Anisimov, and A. Turlikov, "eMBMS LTE Usage to Deliver Mobile Data," in *2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Oct 2014, pp. 60–65.

[8] 3GPP TS36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures," 3GPP, Standard, Jan. 2019.

[9] 3GPP TS36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," 3GPP, Standard, Dec. 2018.

[10] M. Rupp, S. Schwarz, and M. Taranetz, *The Vienna LTE-Advanced Simulators: Up and Downlink, Link and System Level Simulation*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[11] "Evaluation Methodology Document (EMD)," IEEE 80216m, Tech. Rep., July 2008.

[12] S. Schwarz and M. Rupp, "Throughput Maximizing Feedback for MIMO OFDM Based Wireless Communication Systems," in *2011 IEEE 12th International Workshop on Signal Processing Advances in Wireless Communications*, June 2011, pp. 316–320.

[13] C. Liu, C. Shen, J. Chuang, R. A. Rouil, and H. A. Choi, "Evaluating unicast and MBSFN in public safety networks," 2019 forthcoming.

[14] L. Rong, O. B. Haddada, and S. Elayoubi, "Analytical Analysis of the Coverage of a MBSFN OFDMA Network," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, Nov 2008.

[15] 3GPP TS36.212, "Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and Channel Coding," 3GPP, Standard, Jan. 2019.

[16] 3GPP TR25.996, "Spatial Channel Model for Multiple Input Multiple Output (MIMO) Simulations," 3GPP, Standard, July 2017.

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Throughput Analysis between Unicast and MBSFN from Link Level to System Level." Paper presented at 2019 IEEE 90th Vehicular Technology Conference: VTC2019-Fall, Honolulu, HI, US. September 23, 2019 - September 25, 2019.

Distributed Resource Allocation Schemes for Out-of-Coverage D2D Communications

Jian Wang, Richard A. Rouil, and Fernando J. Cintron
Wireless Networks Division, Communications Technology Laboratory
National Institute of Standards and Technology
Emails:{jian.wang, richard.rouil, fernando.cintron}@nist.gov

***Abstract*—In many public safety scenarios, Device-to-Device (D2D) communication should be capable of handling out-of-coverage situations, ensuring that D2D devices can communicate directly without the aid of network infrastructure. While a number of resource allocation schemes for D2D communication in Long-Term Evolution (LTE) have been proposed, the majority consider only in-coverage scenarios that rely on a centralized controller to coordinate among D2D devices, and are unable to handle out-of-coverage communication. To address this issue, in this paper we investigate a set of distributed resource allocation schemes for out-of-coverage D2D group communication. To be specific, we first provide guidelines concerning how to allocate D2D resources based on Modulation and Coding Scheme (MCS), Physical Resource Block (PRB) size, and Time Resource Pattern (TRP) to meet the QoS requirements of applications. We then design three distributed resource allocation schemes that select PRBs in the resource pool and/or adjust the transmitting power based on the level of available information about the network environment. The first scheme, called the basic random allocation scheme, allows the transmitting User Equipment (UE) to randomly select resource blocks from the resource pool and transmit data at maximum power. The second scheme, which enhances the basic random scheme and is denoted as the Received Signal Strength (RSS)-based random allocation scheme, leverages RSS to reduce the power consumption of the transmitting UEs and interference to other groups. The third scheme proposed, designated the interference-aware allocation scheme, allows the transmitting UE to explore the interference experienced by receiving UEs within its D2D group and select resource blocks that interfere with the smallest number of UEs within the group. In doing so, the communication interference among transmitting UEs can be reduced, resulting in a higher probability of system coverage compared with the two random allocation schemes. To evaluate the designed distributed resource allocation schemes, we conduct extensive performance evaluation, validating their effectiveness in a variety of deployment scenarios.**

***Keywords*—D2D Communication, Out-of-Coverage, Distributed Resource Allocation.**

I. Introduction

To carry out public safety response and disaster recovery, emergency response communication is critical [1], [2], [3], [4], [5], [6]. When a large disaster occurs, network infrastructure will likely be overloaded, if not damaged and unavailable. To maintain continuous communications among first responders and victims, technologies to enable direct communication are paramount. Device-to-Device (D2D) communication, as one of several viable technologies to enable direct communication, is considered as a critically important technology in public safety research and development. With D2D, first responders and victims can directly communicate, enabling the sharing and collection of information, and providing situation awareness of public safety events. In these public safety scenarios, D2D communication needs to operate under out-of-coverage conditions, in which D2D devices need to communicate directly without the aid of network infrastructures [7], [8].

To carry out resource allocation in LTE-based D2D out-of-coverage situations, the following challenges need to be addressed. First, communication cannot rely on centralized controllers (base stations, etc.) to conduct resource allocation for D2D UEs. While a number of resource allocation schemes have been proposed, most consider only in-coverage D2D scenarios, where centralized controllers are required to coordinate resources. Additionally, these schemes often assume that complete knowledge of the Channel State Information (CSI) of communication and interference channels are available. Nonetheless, such an assumption is not applicable to the out-of-coverage D2D scenarios investigated in this paper. Second, in D2D group communication, as no physical layer feedback exists, little information regarding CSI is available. Thus, how to leverage the available information to improve system performance becomes a challenging issue. Third, UEs (including both transmitters and receivers) have less computation capabilities and are more sensitive to power consumption than larger and more heavily equipped base stations. Thus, existing complex resource allocation schemes become infeasible on UEs due to the limited computing and energy resources. Thus, the design of lightweight and distributed resource allocation schemes is essential to enable out-of-coverage D2D communications.

To address the issues presented thus far, in this paper we make the following concrete contributions.

- **Problem Formalization.** We formalize the resource allocation problem for out-of-coverage D2D communication in LTE-based networks. As the problem is a multi-dimensional issue, we consider several key factors together, including Modulation and Coding Scheme (MCS), Physical Resource Block (PRB) size, Time Resource Pattern (TRP) and transmitting power, to satisfy the quality of service (QoS) requirements for public safety applications. Once we complete the selection of these key factors, other important decisions include which PRBs in the pre-allocated resource pool should be used and what is the transmitting power level that should be utilized to transmit the signal so that the overall system coverage

Wang, Jian; Rouil, Richard A.; Cintron, Fernando. "Distributed Resource Allocation Schemes for Out-of-Coverage D2D Communications."
Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 13, 2019.

probability can be maximized.

- **Distributed Resource Allocation Schemes.** Based on the formalized problem, we design three distributed resource allocation schemes. Particularly, we first design a basic random allocation scheme, which allows the transmitting UE to randomly select resource blocks from the pre-configured resource pool. This scheme does not assume that the transmitting UE has any knowledge of the network. We then propose the Received Signal Strength (RSS)-based random resource allocation scheme that leverages the available information of the network so that the overall system performance can be improved. The RSS-based random scheme aims to reduce the power consumption of the transmitting UE by leveraging the D2D discovery service and identify the maximum path loss in the D2D group. Finally, we design the interference-aware allocation scheme, which allocates resources by avoiding interference of transmitting UEs. With this scheme, the transmitting UE collects information pertaining to the interference experienced by the receiving UEs in its group, and further selects resource blocks that are used by the least number of UEs. By doing this, the communication interference among transmitting UEs can be reduced so that the system coverage probability can be improved.
- **Extensive Evaluation.** We conduct extensive performance evaluation to show the effectiveness of our proposed schemes in allocating communication resources in out-of-coverage D2D communications with respect to system coverage probability, which is defined as the probability of average received Signal-to-Interference-plus-Noise Ratio (SINR) being larger than the minimum threshold required to successfully decode the information. We also measure the power consumption reduction of the RSS-based random scheme and investigate the tradeoffs of power savings and system coverage probability. Our findings show that the basic random scheme works effectively when the region is not dense (i.e., only a few communication groups exist in the region and the receivers in a group are closely located around the transmitter). When the geographic sizes of groups is small, power control could be used to reduce UE transmission losses with a marginal sacrifice in the coverage performance. When the region becomes more dense, the performance of the two random resource selection schemes deteriorate. Furthermore, the interference-aware scheme can significantly improve the coverage by mitigating interference among groups.

The remainder of this paper is organized as follows: In Section II, we provide a literature review of research relevant to our study. In Section III, we introduce the system model. In Section IV, we introduce the problem formalization, design rationale, and our designed distributed schemes in detail. In Section V, we present the performance evaluation results. Finally, we conclude the paper in Section VI.

II. Related Work

In this section, we review research works relevant to our study. D2D communication is essential for supporting public safety applications, as it enables continuous communication when network infrastructure is either damaged or overloaded [3]. D2D communication has the potential to not only improve communication resilience, but also improve spectrum efficiency, system throughput, energy efficiency, and system delay performance [9], [5], [6].

Generally speaking, D2D communication can be categorized as either in-coverage or out-of-coverage [9], [5], [6]. Particularly, in-coverage D2D means that D2D UEs are within the coverage range of network infrastructure, while out-coverage D2D means that D2D UEs are not in the coverage range of network infrastructure. When UEs are in-coverage, they may share the same radio resources with cellular services (underlay) [10], or use their dedicated resource pool for communication (overlay) [11]. For in-network D2D, traffic can either directly flow between two UEs or flow through the base station. In underlay cases, one challenge is how to minimize the impact of D2D on cellular networks and improve spectrum efficiency, while in overlay cases, challenges include how to choose proper resource pools for D2D so that resource use can be maximized. For out-of-coverage D2D, the primary challenge is how to effectively manage communication resources without the aid of base stations as centralized controllers or coordinators.

There is a body of research on D2D resource management, most of which are concerned with in-coverage D2D [5], [6], [9]. These existing D2D resource management efforts address coverage issues by designing a variety of resource allocation schemes to optimize certain performance metrics, including maximizing system throughput [12], [13], minimizing link outage probability [14], and maximizing overall energy efficiency [15], among others. The optimization objectives are also subject to a number of constraints, including transmission power, minimum QoS requirements, and physical resources, among others. D2D resource allocation also needs to consider a number of other factors, such as user traffic scheduling (i.e., how often the data will be transmitted), MCS selection, the size of physical resource blocks, transmission power, and mode selection to determine whether to utilize D2D or cellular communication modes (again, only for in-network scenarios). All of these factors are interconnected and jointly affect system performance. Thus, to simplify the problem, existing solutions often optimize one or several parameters of the entire parameter set. Commonly used techniques include integer/linear programming, convex optimization, mixed integer nonlinear programming, game theory approaches, and heuristic algorithm design, among others.

Most existing D2D research has focused on in-network underlay scenarios, ranging from the simple setup of one cellular UE and one D2D pair [10], to multiple D2D pairs and multiple cellular UEs [16], [17]. One focus is interference management, which includes issues such as how to minimize the impact of D2D communication on existing cellular services and how to manage cellular-to-D2D interference [16], [11]. For example, Su *et al.* [16] focused on D2D underlay in-band communication, studying how to control the interference from D2D UEs to the primary cellular UE by designing proper resource allocation and mode selection strategies. To solve the problem, a particle-swam optimization (PSO-MSRA)-based scheme was designed to maximize system throughput while guaranteeing the minimal rate requirement for all D2D users. Ye *et al.* adopted a game theory-based scheme and formulated the resource allocation in D2D underlay into a two-stage optimization problem [17]. While the proposed scheme might not find the optimum solution, it allows resources to be allocated in a distributed fashion. Notice that most of existing efforts on in-network are on underlay D2D, and perfect CSI is commonly assumed. There are some existing efforts on in-network overlay D2D as well. For instance, Lee *et al.* [11] studied D2D overlay resource allocation using graph-coloring to maximize the sum rate of D2D links. As a centralized solution, in their scheme, the eNodeB collects a list of high-interfering D2Ds and allocates resource to D2D transmitters.

As our work targets public safety scenarios, communication, and especially group communication, among public safety personnel need be maintained in the event of large-scale emergencies and disasters. Our research focuses on out-of-coverage D2D group communications, which have not been well explored. In our study, we take into account key factors that can be leveraged by a D2D UE locally, and design our solution to be practical, such that it could be implemented within the existing 3GPP framework with no or only slight modifications. We design three distributed schemes to conduct resource allocation for D2D communication to accommodate the special challenges associated with out-of-coverage communications.

III. System Model

In this section, we introduce the system model. In our study, we consider D2D group communication scenarios, in which a number of D2D nodes (i.e., UEs) that belong to different function groups are deployed in a geographical location and perform public safety missions. Within a group, there is one transmitter UE and multiple receiver UEs. The communications between transmitter UE and receiver UEs are performed directly through D2D communication links. Since UEs are not covered by network infrastructure, such as cellular networks, each UE has a pre-configured resource pool with K units in order to perform D2D communications in out-of-coverage scenarios. Without loss of the generality, we assume all UEs are configured with the same resource pool settings.

Figure 1 illustrates the network model, in which the deployment area is a circular area A with radius R. Within A, there are M groups uniformly randomly deployed, denoted as $G_1, G_2, \ldots, G_M$. Within a group G_i ($i \in [1, M]$), there is an active transmitter UE TXi and all other UEs within the same group are uniformly randomly located in a circular region that is centered on the transmitter UE_i with radius r.

Channel gain between transmitter TX_i ($i \in [1, M]$) and receiver UE_j is denoted as g_{ij}. When we compute the channel gain, path loss, large-scale shadowing, and small-scale fading are considered [18]. We also assume that the channel is a slow changing and semi-static. If we can collect coarse channel state information (e.g., CSI) using upper layer signaling (e.g., D2D discovery messages), such information could be used to guide resource allocation.

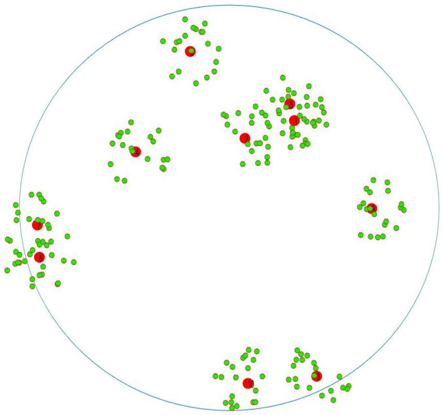


Fig. 1: Network Model

For transmitter TX_i (the transmitter in group i), we denote its transmission power as P_i. We assume all the transmitting groups are running the same application, such as mission critical voice. Resource allocation decisions include the selection of MCS, PRB size, TRP, transmission power, and PRB locations in the resource pool. We first allocate D2D resources by selecting MCS, PRB size, and TRP to meet QoS requirements of the application, such as requirements for throughput and delay. Based on the allocation scheme that we choose, we then divide the resource pool into a number of channels. Each channel occupies the same number of resources and a transmitter is using one channel.

Given M transmitters and K channels for D2D communication, we define the resource usage matrix U as

$$U_{M\times K} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \vdots & & \vdots & & \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix}$$

where $U_{i,j} =$

$$\begin{cases} 1, & \text{if TX } i \text{ uses channel } j \\ 0, & \text{if otherwise} \end{cases} \tag{1}$$

In our study, we consider an outdoor environment and adopt the 3GPP outdoor-to-outdoor (O2O) path loss model [19]. The line-of-sight (LOS) path loss PL_{LOS} and non-line-of-sight (NLOS) path loss PL_{NLOS} are defined as

$$\begin{aligned} PL_{LOS} = {} & 40\log_{10}(d) + 7.56 - 17.3\log_{10}(h_{BS}^{'}) \\ & - 17.3\log_{10}(h_{MS}^{'}) + 2.7\log_{10}(f_c) \\ & + 20\log_{10}(f_C/f_{REF}), \end{aligned} \tag{2}$$

$$\begin{aligned} PL_{NLOS} = {} & (44.9 - 6.55\log_{10}(h_{BS}))\log_{10}(d) + \\ & 5.83\log_{10}(h_{BS}) + 16.33 + 26.16\log_{10}(f_C) - 5. \end{aligned} \tag{3}$$

Here, d is the distance between transmitter and receiver, and $h_{BS}^{'}$ and $h_{MS}^{'}$ are the effective heights for transmitter and receiver in meters, respectively, both of which are set to 0.8 m.

Also, f_c is the carrier frequency, set to 700 MHz, and the reference frequency f_{REF} is 2 GHz. For large-scale shadowing, log-normal shadowing with 7 dB standard deviation is used. For small-scale fading, Rayleigh fading is adopted. The probability of LOS is a function of distance following $P_{LOS} = min(18/d, 1)(1 - exp(-d/36)) + exp(-d/36)$.

IV. Proposed Schemes

In this section, we introduce our schemes in detail. Particularly, we first introduce coverage probability, which is used to evaluate our schemes, and present the optimization of coverage probability to motivate the design of distributed resource allocation. We then outline the problem definition and introduce the design rationale. Finally, we present our proposed schemes to allocate communication resources for out-of-coverage D2D.

A. Coverage Probability

In this study, we use coverage probability as the primary evaluation metric. Generally speaking, coverage probability is defined as the probability that the average received SINR is above the minimum threshold required to successfully decode the transmitted message. The coverage probability of a single transmitter and receiver link is a function of distance, and its expression is [18]

$$\frac{1}{\sqrt{\pi}\Gamma(m)}\int_{-\infty}^{+\infty}\exp(-x^2)\Gamma(m, \frac{m\gamma}{10^{\frac{\sqrt{2}\sigma_s x+\mu}{10}}})\, dx. \tag{4}$$

Here, $x = \frac{10log_{10}w-\mu}{\sqrt{2}\sigma}$ and $w = 10^{(\frac{\sqrt{2}\sigma_s x+\mu}{10})}$, σ_s is the standard deviation of the Log-normal shadowing, and μ is the received power after considering the path loss in dBm. $\Gamma(m)$ is the $\Gamma(.)$ function with an input of m, where $m = 1$ for Rayleigh fading. Also, γ is the decoding threshold that is dependent on the thermal noise, receiver noise figure, and SNR decoding threshold.

Considering interferences, the SINR of UEj in group i using channel k, denoted as $SINR_{j,k}$, can be computed as

$$SINR_{j,k} = \frac{P_{i,k}g_{i,j}}{\sum_{l=1,l\neq i}^{M} U_{l,k}P_{l,k}g_{l,j} + N}, \tag{5}$$

where $P_{i,k}$ is the transmitted power of transmitter UE i using channel k, $g_{i,j}$ is the channel gain between transmitter UE TX_i and receiver UE UE_j, M is the size of transmitter UEs, $U_{l,k}$ is the element of matrix defined in Equation (1), and N is the noise floor on the receiver, which includes thermal noise and noise introduced by the device. Notice that the channel gain includes the path loss, shadowing, and small scale fading.

For a given deployment (fixed transmitter and receiver locations), the path loss and shadowing for each link are fixed, we first compute the UE's average coverage probability over small-scale fading. For UE_j using channel k, its SINR can be written as

$$SINR_{j,k} = \gamma = \frac{g_0\omega_0}{\sum_{l=1,l\neq i}^{M} U_{l,k}g_l\omega_l + N}, \tag{6}$$

where ω_0 and ω_l are the local mean powers of the desired signal S_0 and interference signal S_l, i.e., the received powers after considering path loss and shadowing, which are fixed for a deployment. Also, g_0 and g_l are the power gain of the Rayleigh fading.

We define the coverage probability of UE_i as

$$\begin{aligned} P(\gamma \geq \beta|\omega) &= P(\frac{g_0\omega_0}{N + \sum_{l=1,l\neq i}^{M} U_{l,k}g_l\omega_l} \geq \beta|\omega), \\ &= P(\beta^{-1}g_0\omega_0 - \sum_{l=1,l\neq i}^{M} U_{l,k}g_l\omega_l \geq N|\omega), \\ &= P(S \geq \sum_{l=1,l\neq i}^{M} U_{l,k}g_l\omega_l + N|\omega). \end{aligned} \tag{7}$$

Here, β is the SNR threshold for successfully decoding the information from noisy communication channels in a given probability. Since g_0 and g_l are the power gains of the Rayleigh fading, the probability density function (PDF) of $S = \beta^{-1}g_0\omega_0$ is $f_S(s) = \frac{\beta}{\omega_0}\exp(\frac{-\beta s}{\omega_0})$ and let $y_l = U_{l,k}g_l\omega_l$, Equation (7) can be rewritten as:

$$\begin{aligned} P(\gamma \geq \beta|\omega) &= \int \cdots \int_Y \int_{N+\sum_{l=1,l\neq i}^{M} y_l}^{\infty} f_S(s)f_Y(y_1, \ldots, y_M)\, dsdY \\ &= \exp(-\beta_0 N)\int \cdots \int_Y exp(-\beta_0 \sum_{l=1,l\neq i}^{M} y_l) \\ &\quad f_Y(y_1, \ldots, y_M)\, dY \end{aligned} \tag{8}$$

Here, $\beta_0 = \frac{\beta}{\omega_0}$.

Since the UEs are deployed independently, y_i is an independent and identically distributed (i.i.d) random variable. Thus, we have

$$P(\gamma \geq \beta|\omega) = exp(-\beta_0 N)\prod_{l=1,l\neq i}^{M}\int_0^{\infty} exp(-\beta_0 y_l)f_Y(y_l)dy_l. \tag{9}$$

If the transmitter picks a channel randomly from a pre-allocated pool, we have PDF of $y_l = U_{l,k}g_l w_l$ as,

$$f_Y(y_l) = (1 - p_i)\delta(y_l) + p_i * \frac{1}{\omega_l}\exp(-\frac{y_l}{\omega_l}), \tag{10}$$

where $p_i = \frac{1}{N_{ch}}$ and N_{ch} is the total channel numbers.

After plugging in Equation (10), we have

$$\begin{aligned} P(\gamma \geq \beta|\omega) &= exp(-\beta_0 N)\prod_{l=1,l\neq i}^{M}\int_0^{\infty} exp(-\beta_0 y_l)((1 - \frac{1}{N_{ch}})\delta(y_l) \\ &\quad + \frac{1}{N_{ch}}\omega_l exp(-\frac{y_l}{\omega_l}))dy_l, \\ &= exp(-\beta_0 N)\prod_{l=1,l\neq i}^{M}\frac{1 + \beta_0\omega_l(1 - \frac{1}{N_{ch}})}{1 + \beta_0\omega_l}. \end{aligned} \tag{11}$$

Once we have the coverage probability for each UE with Equation (11), we can obtain the average UE's coverage probability in a fixed deployment. We then simulate the coverage probability over different deployments to obtain the average coverage probability for given system configurations, which vary by region size, number of groups, group region size, and others.

In a multi-group environment, we can choose matrix U such that the UE's average coverage probability can be maximized, i.e.

$$\begin{aligned} \max_{R} \sum_{i,k} P(S_{i,k} \geq t^{-1} + \sum_{l=1,l\neq i}^{M} U_{l,k} g_l w_l | w) \\ s.t. \sum_{j} U_{i,j} = 1,\ P_i \leq P_{max} \end{aligned} \tag{12}$$

Since the objective function is not a closed formula, directly finding for U matrix is very challenging. In addition, in out-of-coverage scenarios, without physical layer feedback and centralized control, it is difficult to collect sufficient information necessary to obtain the settings to maximize the average coverage probability. Thus, this motivates us to design distributed resource allocation schemes that are feasible in out-of-coverage D2D communications. In our study, we design an interference-aware scheduling scheme to select matrix U. Using the random channel selection scheme as a baseline, we investigate whether the coverage performance can be improved by selecting channels based on detected interference information. Once we have the channel allocation matrix U designed, the average coverage probability of UE i for a fixed deployment can be evaluated as:

$$\begin{aligned} P(\gamma_i \geq \beta|\omega) &= P(S \geq \sum_{l=1,l\neq i}^{M} U_{l,k} t_l + \tau^{-1}|\omega) \\ &= \int_T \cdots \int \int_{\tau^{-1}+\sum_{l=1,l\neq i}^{M} U_{l,k} t_l}^{\infty} f_S(s) \\ & f_T(t_1, \ldots, t_M)\, ds dT, \\ &= \exp(-\beta_0 N) \int_T \cdots \int exp(-\beta_0 \sum_{l=1}^{M} U_{l,k} t_l) \\ & f_T(t_1) \ldots f_T(t_M)\, dT, \\ &= \exp(-\beta_0 N) \prod_{l=1,l\neq i,U_{l,k}=1}^{M} \frac{1}{1+\beta_0 \omega_l}. \end{aligned} \tag{13}$$

Here, $f_T(t_l = g_l \omega_l) = \frac{1}{\omega_l} \exp(\frac{-t_l}{\omega_l})$.

Using Equations (11) and (13), we average out the small-scale fading, which can significantly reduce simulation time. To validate these two equations, we ran Monto Carlo simulations to simulate small-scale fading, and the comparison can be seen in Figure 2 and Figure 3. From both figures, we observe that the full-scale Monto Carlo simulations match well with the analytical results provided by Equations (10) and (13).

B. *Design Rationale*

Recall that, unlike in-coverage communication, in the out-of-coverage case, there is no base station acting as a central controller to allocate resources for each UE, and UEs need to schedule their resources autonomously. Thus, the problem we seek to address is, in the group communication scenario as described in Section III, how can transmitter UEs select resources with consideration for MCS, PRB size, and TRP to satisfy the QoS requirements of D2D communication, and how can UEs select transmitting power and the location of PRBs in the resource pool to improve system coverage probability. Notice that our focus is to improve the coverage probability so that the reliability requirements of public safety applications can be fulfilled. Other objectives (power consumption, system throughput, etc.) can be considered and extended in future extensions to this study.

According to the definition of coverage probability, UE j has coverage if $SINR_j > \beta$. Here, β is the SNR threshold in dB for UEj in order to achieve 10^{-2} Block Error Rate (BLER) after 4^{th} D2D transmissions. Notice that β is MCS dependent.

Depending on the availability of information on the deployment environment, we propose the following resource allocation schemes:

- *Basic Random Allocation Scheme*. In this scheme, each UE randomly selects resources from the resource pool in a uniform fashion based solely on its throughput requirement, without leveraging any knowledge, such as CSI. In this scheme, we do not conduct power control, such that the UE uses the maximum transmission power to send messages (what we consider the default configuration).
- *Received Signal Strength (RSS)-Based Random Allocation Scheme.* In this scheme, we intend to reduce UE power consumption by leveraging the D2D discovery service to identify the maximum path loss in the D2D group. Based on the maximum path loss in a group, the transmission power can be adjusted in the transmitting UEs, leading to reduction in power consumption, and potentially reduced interference with neighboring groups. Notice that UEs still randomly select the resources from the resource pool in a uniform fashion.
- *Interference-Aware Allocation Scheme.* In this scheme, the transmitter UE collects information on the interference experienced by the receiver UEs in its group, and selects the channels that can be detected by the least number of UEs in its group. By doing this, the communication interference among UEs can be reduced, such that the overall system coverage probability can be improved.

In the following, we describe our designed resource allocation schemes in detail.

C. *Basic Random Allocation Scheme*

In the basic random allocation scheme, to satisfy the throughput requirements of UEs, we need to select the proper combination of MCS, PRB size, and TRP. Notice that the resource pool size and transmission period length are pre-configured for all UEs to communicate outside cellular coverage. To illustrate this resource allocation scheme, we give an example. To support AMR-WB (Adaptive Multi-Rate - Wideband Speech Codec) voice applications, a throughput of 12.65 kbits/s is required, which means 253 bits for every 20 ms. After adding 3 bytes of RoHC (Robust Header Compression), LLC (Logical Link Control), and MAC (Media Access Control) headers, a transport block with a minimum size of 300 bits is required. If the period length is 40 ms, and 8

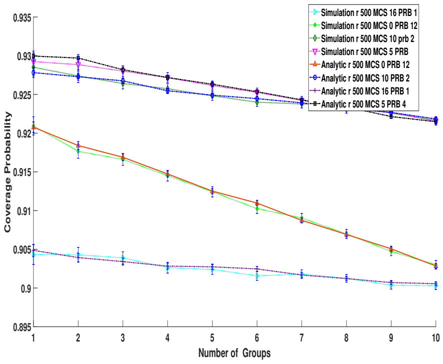


Fig. 2: Rayleigh Fading Simulation vs. Analytical (Eqn. (11)) for Random Channel Selection

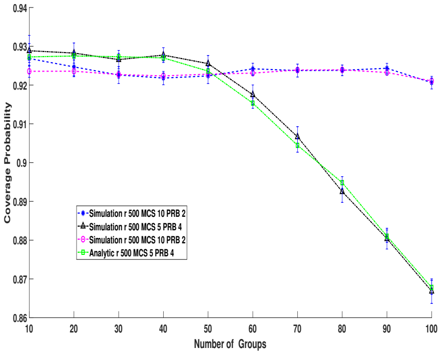


Fig. 3: Rayleigh Fading Simulation vs. Analytical (Eqn. (13)) for Interference Aware Scheme

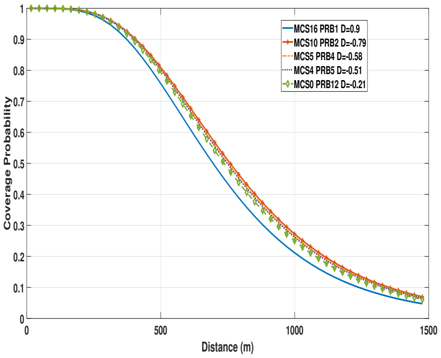


Fig. 4: Coverage Probability Comparison

MCS	PRB	Threshold (dB)	10$log_{10}PRB$	D
16	1	0.9	0	0.9
10	2	-3.8	3.01	-0.79
4	5	-7.5	6.99	-0.51
0	12	-11	10.79	-0.21
5	4	-6.6	6.02	-0.58

TABLE I: D values for different MCS and PRB combinations

subframes within the period are used for transmission, considering 4 retransmissions, a new transport block is transmitted every 20 ms.

Once the number of subframes to be transmitted in a transmission period is determined, we need to determine the transport block size to meet the QoS requirement (i.e., the amount of data to be transmitted in a subframe) of UEs. To satisfy the designated transport block size, we have several MCS and PRB size combinations. The problem is how to select the desired MCS and PRB pair among these combinations. As shown in our prior study [18], to successfully decode a transmitted message through a D2D link, a higher MCS value requires a higher SNR threshold. In contrast, if we choose a low MCS, we will use more PRBs, and noise floor will rise for the channel. The received signal strength should be greater than the receiver sensitivity to decode the signal successfully, and thus the maximum path loss that we can tolerate to achieve reliable communication is $P_{TX} - N - NF - SNR_{th}$ [18]. Here, N is the thermal noise floor and depends on the channel bandwidth, NF is the device noise figure, and SNR_{th} is the SNR to achieve 1 % BLER after the 4^{th} transmission.

Thus, with fixed transmitting power and noise figure, maximizing the path loss is equivalent to minimizing the sum of N and SNR_{th}. To maximize the propagation distance (i.e., maximizing the path loss that the receiver can tolerate), we consider the following objective: for each MCS and PRB size combination, we compute the utility function D_{max} as follows:

$$D_{max} = 10 \log_{10} PRB + SNR_{th}. \tag{14}$$

We then pick the MCS and PRB size combination so that the smallest D value can be realized. We thus search through the MCS and PRB combinations that meet the throughput requirements. Table I illustrates some examples of D values. From the table, we can see that, by choosing the minimum D value, MCS 5 and PRB size 4 will be selected.

The average coverage probability of a link between one transmitter and one receiver for Rayleigh fading channel is computed by leveraging our prior work [18], and the results are shown in Figure 4. In the figure, we show the coverage probability for 5 different PRB and MCS combinations. Among these combinations, as a combination of MCS 16 and PRB 1 leads to the highest D value, the worst performance is achieved, while as a combination of MCS 10 and PRB 2 leads to the lowest D value, the best performance is achieved. A combination of MCS 5 and PRB 4 leads to the second best, resulting in the performance being close to the best. For the single D2D link, by picking the MCS and PRB combination with the lowest D value, we can maximize its coverage probability.

Notice that the aforementioned result is only for D2D links without interference from other transmitters. If each D2D transmitter uses this strategy, in a multi-group environment, it may not achieve the overall best system performance with respect to coverage probability. The resource allocation scheme using the smaller PRB can cause less interference to neighboring groups, which could outperform the best pair (e.g., MCS 5, PRB 4) with the increase of D2D groups in the region. Thus, depending on how dense the area is (i.e., the number of D2D groups packed in this area), a UE needs to give weight to a particular PRB size in order to achieve better system performance. With the pre-knowledge of the number of D2D groups in the deployed area, the UE can select the MCS and PRB pair accordingly so that the overall system performance can be improved. Algorithm 1 shows the procedure for selecting the MCS and PRB combination with the smallest D value in order to maximize coverage probability. As Algorithm 1 computes the MCS and PRB pair with the smallest D value by enumerating all combinations, its complexity is $O(n_{MCS} * n_{PRB})$, where n_{MCS} is the total number of MCSs and n_{PRB} is the total number of PRBs.

Algorithm 1: Algorithm for Selecting MCS and PRB Combination

```
Input: Throughput requirement and TRP settings
Output: The selected MCS and corresponding PRB
Compute minimal Transport Block (TB) Size =
  Throughput (bits/s)/Number of subframes transmitted
  Per Second
Check Table 7.1.7.2.1-1 [20] and list all the MCS and
  PRB combinations, whose corresponding TB sizes
  are equal to or above the minimal TB size
Compute utility function value $D$ using Equation (14)
return The MCS and PRB combination that has the
  smallest $D$ value
```

In the following, we introduce two other schemes that leverage available network information to improve the system performance.

D. RSS-based Random Scheme

We now introduce the RSS-based random resource allocation scheme, which adopts power control to reduce the energy consumption of the transmitter UEs and potentially reduces interference to neighboring groups. In the RSS-based random scheme, we leverage the D2D discovery service. There are two discovery modes: (i) *Model* A, and (ii) *Model* B. In Model A, the UE sends a discovery message autonomously, while in Model B, the UE is polled by neighboring nodes and sends the discovery response. Through Model B, the transmitter UE can send inquiries to its D2D group. Based on the SD-RSRP (Sidelink Discovery Reference Single Received Power) of the discovery response message, the path loss of UEs in the group can be estimated. Notice that RSRP is the average power received on the resource elements that carry the Demodulation Reference Signal (DMRS) of a decoded PSDCH (Physical Sidelink Discovery CHannel) signal. From the SD-RSRP and UE's transmitter power, the time averaged channel loss from the receiving UE in the group to the transmitter UE can be estimated. As D2D uses the LTE uplink spectrum for communication, D2D channels in both directions are reciprocal. Thus, we can estimate the channel loss from the transmitter UE to the receiver UE based on channel reciprocity.

Once the maximum channel loss in a group is available, the transmitter UE can use the same MCS and PRB as identified using the basic random allocation scheme, but instead of always transmitting using the maximum power, it may transmit with a reduced transmission power using the channel loss information. We assume network deployment is not fast-changing, and thus the CSI collected in the discovery process can be used to assist communication.

Power control can be conducted either open-loop or closed-loop. The open-loop control is to set the transmission power based on path loss and shadowing information, while the closed-loop control is used more often to accommodate the fast-fading effect. By controlling the transmitting power, UEs can not only save energy and improve battery life, but also reduce interference with other transmitting UEs in neighboring groups while preserving QoS of the UEs. Since D2D group communications do not have physical layer feedback and only coarse channel state information is available, we can only estimate path loss and shadowing effects from the discovery message from its group UEs, and the channel information is used to set transmitting power.

In this scheme, we introduce the compensation factor (CF), which denotes how much compensation we want for the channel loss (including both path loss and shadowing). The transmitted power after power control is

$$P_{tx} = \min((1-CF)P_{max} + CF \times PL_{max} + \text{noise floor}, P_{max}), \quad (15)$$

where the PL_{max} is the maximum channel loss of the radio link between transmitter UE and its group UEs. When CF is 1, P_{tx} is the minimum transmitting power to bring the received average power just above the noise floor. With the growth of CF, we can increase the transmitting power to have more margin and account for small-scale fading and the interference from other groups. Protocol 1 shows the detailed procedure for conducting power control.

Protocol 1 RSS-based Random Scheme

Inputs: Preselected MCS and PRB, and CF value
Output: UE transmitting power
Protocol:

1) Transmit UE sends discovery request using maximum power
2) UEs send back discovery responses after decoding the request using the maximum power
3) Transmit UE calculates the maximum channel loss experienced by its group and sets its transmit power using Equation (15)

Protocol 2 Interference Aware-Based Scheme

Inputs: Preselected MCS and PRB, and resource pool (channels)
Output: Which PRBs to select in the resource pool
Protocol:

1) Each UE monitors the channel that it can detect, and builds a list of channel with RSS greater than -105 dBm
2) Transmit UE sends discovery request using maximum power
3) UEs send back discovery responses with the channel list using the maximum power after decoding the request
4) Transmit UE sorts the interference channels by interference level, i.e. the number of group UEs that can detect this channel
5) Compute the channel that has the lowest level of interference

E. *Interference-Aware Allocation Scheme*

The resource allocation of the two aforementioned schemes are based on the random selection strategy, meaning that the transmitter UE randomly selects the PRB location uniformly from the resource pool. Notice that, in our study, we assume all transmitting UEs use just one channel, and each channel contains the same number of resource blocks. Thus, the problem becomes how to pick the channel for transmitter UEs. If a transmitter UE selects a channel randomly without knowledge of its environment, it may not achieve desirable system performance.

To consider interference, the transmitter UE first collects information about the interference experienced by the receiver UEs within its D2D group. This can be carried out by sending a discovery query message to nearby UEs. The UEs within the same group, after receiving the message, then piggy back the channels that it can detect (i.e., average RSS is above -105 dBm) using the discovery response message. When the information is collected, the transmitting UE will rank each channel by the number of group UEs that can detect that channel (i.e., the number of group members that are interfered by that channel). The channel detected by the least number of transmitters will be selected by the transmitter UE. If there are multiple channels that could satisfy the requirements, we will select the channel that has been used most recently. If no such information available, we will randomly pick one from the multiple channels. For example, if two transmitters can be detected by the same receiver, these two transmitters become interferers to each other, and we should try to put these two on different channels so that interference between them can be avoided. Thus, if a transmitter's total number of interference channels is less than the size of the channel pool, the transmitter can use one of the unused channels to transmit. If the transmitter is located within a much more dense area, and all the channels have been used by all neighbor transmitters, the channel detected by the minimum number of UEs in its group will be picked. Protocol 2 shows the detailed procedure for conducting the channel selection with the least interference.

V. Performance Evaluation

In this section, we present our performance evaluation. In particular, we first introduce the evaluation methodology and provide evaluation results.

A. *Methodology*

To evaluate the performance of the D2D group communication system, we consider coverage probability, defined in Section IV, as the key metric. To evaluate the power saving performance, we measure the power saved in the transmitting UE as the ratio of the power usage from all transmitter UEs transmitting at maximum power to the power usage when the power control scheme is in place, the results of which are presented in units of dB.

To simulate a group communication scenario, a number of UE groups are deployed in a circular region with radius 3000 m. Within each group, 20 receiver UEs are deployed within a small circle around the transmitter UE. We evaluate the performance by varying the number of groups in the region in order to simulate a region with different levels of density. We also evaluate the impact of group size on performance by varying the closeness of the receivers to the transmitters. We assume all the transmitter UEs have full buffers so that the transmission is continuous. For fixed locations of transmitters and receivers, we generate log-normal random variables to simulate the shadowing effect of the channel between each transmitter and receiver pair. With the deployment and shadowing information, we can compute the area mean power of a receiver and use Equations (10) and (13) in Section IV-A to compute the average coverage probability of a deployment, and we can simulate hundreds of deployments to obtain average performance.

B. *Results*

1) Basic Random Allocation Scheme: Figure 5 illustrates the coverage probability vs. the number of D2D groups in a range of [1, 10]. Figure 6 illustrates the coverage probability vs. the number D2D groups in a range of [10, 100]. From both figures, we can make the following observations. First, when the number of groups increases, the coverage probability reduces due to the increase in interference. When UEs in a group are geographically close to each other (i.e., r is 300 m), receiving UEs have higher desired signal power and experience less interference from other transmitter groups, such that the average coverage probability is better than for UEs in more dispersed groups (i.e., r is 500 m). For just a single D2D group, the combination of MCS 5 and PRB 4 achieves the best coverage with a small margin over MCS 10 and RPB 2 (due to the smallest D value). With the increase in number of D2D groups, MCS 10 and PRB 2 soon outperform MCS 5 and PRB 4.

Additionally, with PRB 12, since it uses MCS 0, which requires the lowest SNR to decode the signal, when there is only 1 group, it performs better than MCS 16 and PRB 1. However, when the number of UE groups in a region reaches a certain level, the interference becomes a problem. When this

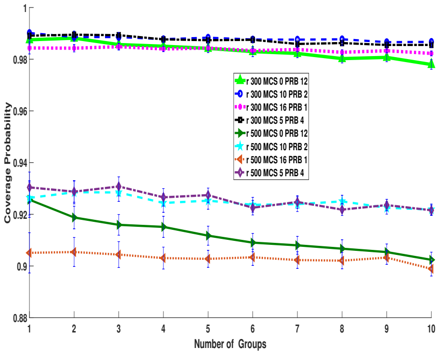


Fig. 5: Basic Random Scheme (R = 3000 m, r = 100 m, 500 m, Group=[1,10])

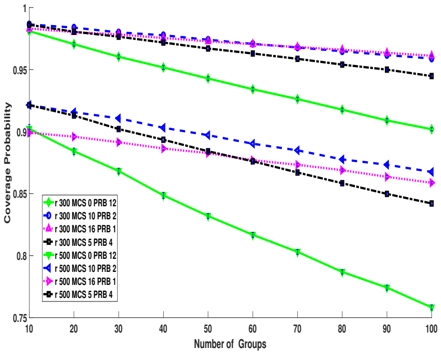


Fig. 6: Basic Random Scheme (R = 3000 m, r =300 m, 500 m, Group=[10,100])

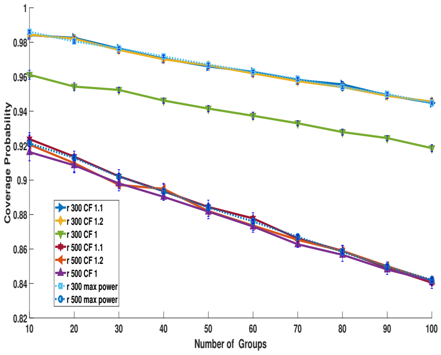


Fig. 7: Coverage Probability of RSS-based Random Scheme (MCS=10, PRB=2)

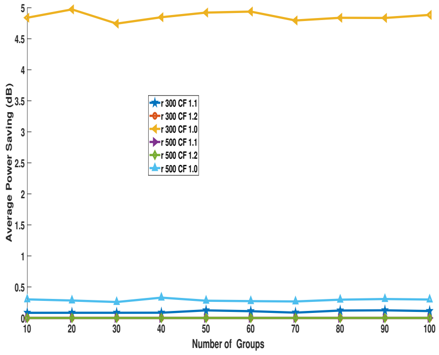


Fig. 8: Power Saving of RSS-Based Random Scheme (MCS = 10, PRB = 2)

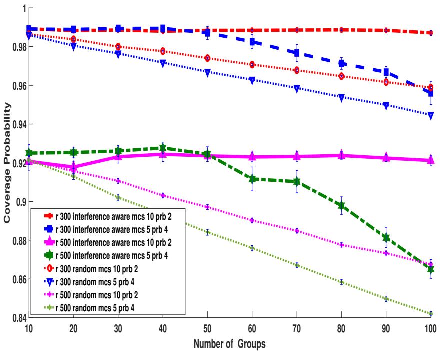


Fig. 9: Coverage Probability of Interference Aware Allocation Scheme (R = 3000 m, r = 300 m, 500 m)

occurs, MCS 12 starts having better coverage probability than MCS 0. This is because, with the fixed resource block size, more channels can be divided for PRB 12 than for PRB 1, leading to a lower chance of collision in the frequency domain. In general, when the system is dense, or more transmitters are deployed in a given region, the resource allocation schemes that using less resources have a performance advantage. Notice that, in the basic random scheme, we do not take into account the interference from different groups. This is the decision made by the transmitter UE itself. If each D2D UE uses this scheme, it may not achieve the overall best average coverage probability. Thus, depending on how dense the area is (i.e., how many D2D groups are packed into the deployment area), the UE more heavily weights a smaller PRB size in order to achieve better coverage.

2) RSS-Based Random Allocation Scheme: Figure 7 illustrates the average coverage probability vs. the number of groups (which varies from 10 to 100), where power control is used with different values of CF. Notice that, when CF is 1, we have enough transmitting power to compensate for the average channel loss due to path loss and shadowing. When CF is greater than 1, we overcompensate for the average channel loss to accommodate for the small-scale fading and the interference from other transmitters. Figure 8 illustrates the average power savings in transmitter UEs as the number of groups varies from 10 to 100, and the system implements the RSS-based random scheme. For both figures, the PRB size is set to 2 and the MCS is fixed at 10.

From these two figures, we can observe that, when r is small (r is 300 m) and CF is 1, UEs can save around 5 dB in transmitting power on average, with about a 2 % drop in the coverage probability. However, when r becomes large (r is 500 m), to compensate for the largest channel loss in a D2D group, the transmitting power is close to the maximum transmitting power and the power savings become insignificant (around 0.3 dB). As a consequence, the coverage probability is comparable to that of the maximum transmitting power. When CF is 1.1 or 1.2, there is almost no power savings, meaning that the transmitting UEs are approximately using the maximum transmitting power. Based on our evaluation results, we confirm that there are tradeoffs between coverage probability and power saving improvements, meaning that greater power savings results in a sacrifice in the form of lower coverage probability.

3) Interference-Aware Allocation Scheme: Figure 9 shows the average coverage probability for two different UE groups with receive UE deployment distances of r=300 m and r = 500 m. We also demonstrate the performance of two MCS configurations (i.e., MCS 10 and MCS 5). We can observe that, when the UEs are closer together, the coverage performance is good, even with a random channel selection. When MCS is 10, since it uses less PRBs, radio resources can be divided into more channels. With the interference-aware scheme, when there are up to 100 groups in the circular region of radius of 3000 m, the interference can be mitigated well and the average coverage probability remains flat. However, when MCS is 5, since it uses double the size of PRB compared to MCS 10, the number of channel divisions that it can allocate is only half. Thus, we can observe that the coverage probability for MCS 5 starts dropping when the number of groups reaches approximately 50, since there are not enough channels to avoid interference. Thus, using neighboring D2D group information, we confirm that coverage probability can be significantly improved, especially when region is dense, assuming appropriate channel selection.

VI. Final Remarks

In this paper, we have addressed the resource allocation issue of out-of-coverage D2D scenarios for public safety communications and investigated three distributed resource allocation schemes. To be specific, we first showed how UEs could schedule resources based on the throughput requirements of applications and maximize transmission coverage. We then investigated how to select the physical block size in the resource pool once the decision is made on MCS and PRB size. To do so, we proposed three distributed resource allocation schemes. In the basic random allocation scheme, no D2D deployment information is relied upon to select resources. In the RSS-based random allocation scheme, power control is used to reduce the energy consumption of the transmitting UEs. In addition to the random allocation schemes, we propose an inference-aware scheme that allows transmitter UEs to collect interference information from nearby D2D groups. By using such information, the interference among UEs can be reduced, leading to performance improvement. We conducted an extensive performance evaluation to validate the effectiveness of our proposed schemes. Our findings show that the basic random allocation scheme works effectively when the region is not dense, the power control scheme is capable of reducing UE transmission power with only a small sacrifice in coverage performance when receiver UEs belong to a group are deployed close to the transmitter, and the interference-aware scheme can significantly improve coverage by mitigating interference among groups.

References

[1] R. A. Rouil, A. I. Manzanares, M. R. Souryal, C. A. Gentile, D. W. Griffith, and N. T. Golmie, "Modeling a nationwide public safety broadband network," *IEEE Transactions on Vehicular Technology*, vol. 8, no. 2, pp. 83–91, 2013.

[2] G. Baldini, S. Karanasios, D. Allen, and F. Vergari, "Survey of wireless communication technologies for public safety," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 2, pp. 619–641, 2014.

[3] G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu, and N. Brahmi, "Device-to-device communications for national security and public safety," *IEEE Access*, vol. 2, pp. 1510 – 1520, 2014.

[4] A. Kumbhar, F. Koohifar, . Gven, and B. Mueller, "A survey on legacy and emerging technologies for public safety communications," *IEEE Communications Surveys Tutorials*, vol. 19, no. 1, pp. 97–124, Firstquarter 2017.

[5] W. Yu, H. Xu, J. Nguyen, E. Blasch, A. Hematian, and W. Gao, "Survey of public safety communications: User-side and network-side solutions and future directions," *IEEE Access*, vol. 6, pp. 70 397–70 425, 2018.

[6] A. Jarwan, A. Sabbah, M. Ibnkahla, and O. Issa, "LTE-based public safety networks: A survey," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2019.

[7] S. Lien, C. Chien, F. Tseng, and T. Ho, "3GPP device-to-device communications for beyond 4G cellular networks," *IEEE Communications Magazine*, vol. 54, no. 3, pp. 29–35, March 2016.

[8] S. Gamboa, F. J. Cintron, D. Griffith, and R. Rouil, "Adaptive synchronization reference selection for out-of-coverage proximity services," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–7.

[9] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.

[10] Y. Pei and Y. Liang, "Resource allocation for device-to-device communication overlaying two-way cellular networks," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013, pp. 3346–3351.

[11] C. Lee, S. Oh, and J. Shin, "Resource allocation for device-to-device communications based on graph-coloring," in *2015 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Nov 2015.

[12] T. Peng, Q. Lu, H. Wang, S. Xu, and W. Wang, "Interference avoidance mechanisms in the hybrid cellular and device-to-device systems," in *2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep. 2009, pp. 617–621.

[13] S. Xu, H. Wang, T. Chen, Q. Huang, and T. Peng, "Effective interference cancellation scheme for device-to-device communication underlaying cellular networks," in *2010 IEEE 72nd Vehicular Technology Conference - Fall*, Sep. 2010, pp. 1–5.

[14] K. Vanganuru, S. Ferrante, and G. Sternberg, "System capacity and coverage of a cellular network with D2D mobile relays," in *MILCOM 2012 - 2012 IEEE Military Communications Conference*, Oct 2012, pp. 1–6.

[15] M. Jung, K. Hwang, and S. Choi, "Joint mode selection and power allocation scheme for power-efficient device-to-device (D2D) communication," in *2012 IEEE 75th Vehicular Technology Conference (VTC Spring)*, May 2012, pp. 1–5.

[16] L. Su, Y. Ji, P. Wang, and F. Liu, "Resource allocation using particle swarm optimization for D2D communication underlay of cellular networks," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*, April 2013.

[17] Q. Ye, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "Distributed resource allocation in device-to-device enhanced cellular networks," *IEEE Transactions on Communications*, vol. 63, no. 2, pp. 441–454, Feb 2015.

[18] J. Wang and R. Rouil, "Assessing coverage and throughput for D2D communication," in *IEEE International Conference on Communication (ICC)*, May 2018.

[19] 3GPP, "Technical Specification Group Radio Access Network; Study on LTE Device to Device Proximity Services; Radio Aspects v.12.0.1 ," 3rd Generation Partnership Project (3GPP), TR 36.843, 2014.

[20] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures,; v.13.2.0," 3rd Generation Partnership Project (3GPP), TS 36.213, 2016.

Multipath Mitigation from FM-AM Correlation

Archita Hati and Craig W. Nelson
Time and Frequency Division
National Institute of Standards and Technology
Boulder, CO/USA
archita.hati@nist.gov

Abstract— **Multipath is one of the limiting factors for an accurate outdoor and indoor localization. We proposed an approach that uses a multipath mitigation derived from the undesired frequency-to-amplitude conversion of a frequency modulated (FM) signal experiencing multipath.**

Keywords*— *fading; frequency modulation; indoor location; multipath

I. Introduction

Multipath propagation is one of the limiting factors for an accurate outdoor and indoor localization [1, 2]. It introduces multipath fading due to the interference of the direct line-of-sight (LOS) signal with reflected signals from objects such as buildings, ground, trees, and other obstacles. The destructive interference of LOS and the reflected signals can create frequency dependent spectral nulls. Indoors, multipath fading occurs frequently and makes it quite difficult to accurately estimate the direct path length. Different multipath compensation schemes have been proposed and implemented over the years to tackle the effect of multipath propagation for accurate location [3, 4]. In this paper, we describe a method of multipath correction that uses correlation between frequency modulation (FM) and amplitude modulation (AM) signals due to frequency-to-amplitude conversion of a FM signal experiencing multipath. The simulation result of the proposed multipath mitigation method is presented.

II. Methods/Results

The proposed approach of multipath mitigation utilizes the undesired frequency-to-amplitude conversion of a frequency modulated (FM) signal experiencing multipath. The basic concept is to measure correlation between the demodulated FM and AM of the received signal under multipath environment and create a control signal for feed-forward correction. One advantage of measuring AM is its simplicity, it requires an AM detector which is a simpler, smaller, and less expensive device.

The block diagram of the scheme used for reducing multipath effects on FM signals is shown in Fig. 1. We implemented this configuration in a LabVIEW simulation. First, the TX signal was generated by frequency modulating a carrier with white Gaussian noise. A model for a simple multipath channel was created by adding a single delayed, attenuated version of the TX signal to itself.

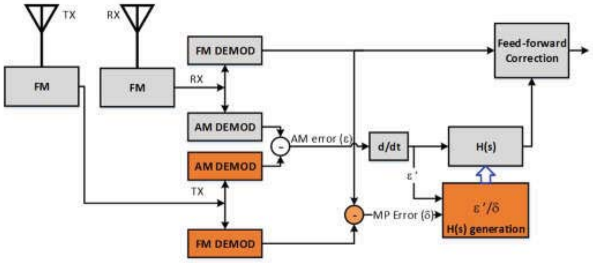


Fig. 1. Block diagram of the FM-AM correlation measurement and compensation technique. The orange blocks are required only for H(s) generation. H(s) is the control transfer function that maps amplitude fluctuations to multipath (MP) error. In case of two-way ranging known modulation symbols can be used for H(s) generation.

Sending the TX signal through this multipath channel generates the received signal RX. The received signal is simultaneously FM and AM demodulated. In addition, the TX signal is also FM and AM demodulated at the transmitter end. The instantaneous multipath error, $\delta(t)$ is determined by subtracting the received FM demodulated signal from the transmitted FM demodulated signal. Similarly, the AM error, $\varepsilon(t)$ is obtained by subtracting the received and transmitted AM demodulated signals. The cross-power spectral density (CPSD) which is a measure of correlation between two time- series, is calculated between $\delta(t)$ and $\varepsilon(t)$ and is given by

$$S_{\delta\varepsilon}\left(f\right)=\frac{2}{T}\left\langle\Delta\left(f\right)\mathrm{A}^{*}\left(f\right)\right\rangle_{m}, \qquad (1)$$

where, $\Delta(f)$ and $\mathrm{A}(f)$ are the Fourier transforms of $\delta(t)$ and $\varepsilon(t)$ respectively, T is the measurement time normalizing the power spectral density (PSD) to 1 Hz, “*” indicates the complex conjugate, and $\langle\ \rangle_m$ denotes an ensemble of m averages. The simulation results in decibels (dB) of $S_{\delta\varepsilon}$ along with S_δ (= PSD of $\delta(t)$) and S_ε (= PSD of $\varepsilon(t)$) are shown in Fig. 2. The degree of correlation between S_δ and S_ε can be described by a correlation function, ρ [5]

$$\rho=\frac{S_{\delta\varepsilon}}{\sqrt{S_{\delta}S_{\varepsilon}}}, \qquad (2)$$

Hati, Archita; Nelson, Craig. "Multipath Mitigation from FM-AM Correlation." Paper presented at 2019 IEEE IFCS-EFTF, Orlando, FL, US. April 14, 2019 - April 18, 2019.

where $\sqrt{S_\delta S_\varepsilon}$ is the geometric mean of S_φ and S_α. The values of ρ range from 0 to 1 and $\rho = 1$ represents 100 % correlation. Fig. 2 shows that the cross-spectrum is almost the expected geometric mean between 10 kHz and 1 MHz offset frequencies indicating almost a 100 % correlation. Further, it can be seen from Fig. 2 that the slope of S_ε is f^0 and S_δ is f^2, so if we generate a control signal utilizing S_ε that is of same magnitude, the same noise slope, and opposite phase as the S_δ , then this control signal can be used in a feedforward approach to reduce the error due to multipath.

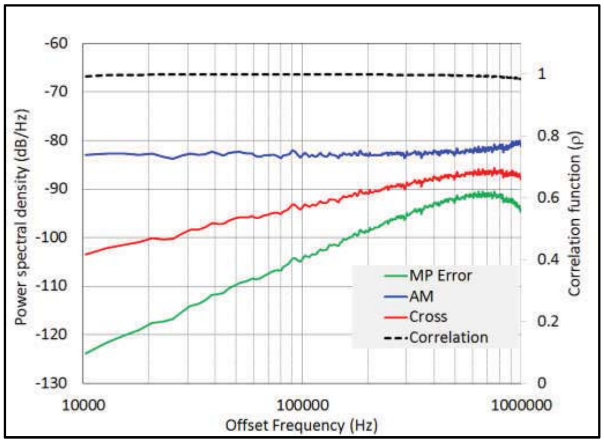
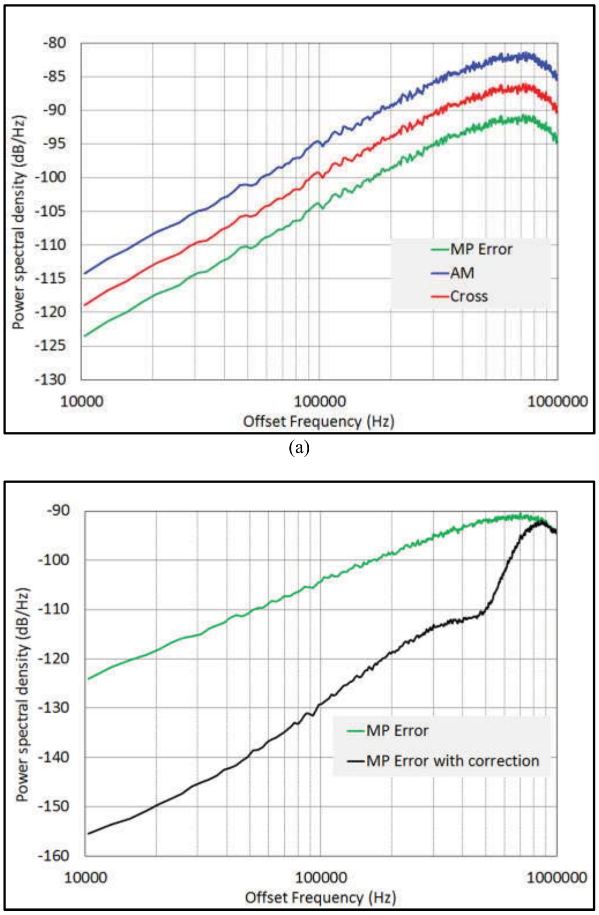


Fig. 2. Plot of the power spectral density $S_{\delta\varepsilon}$ (cross) along with S_δ (multipath error) and S_ε (AM) (left axis). The plot shows almost 100 % correlation (ρ = 1) as shown on the right axis. For this simulation, a 100 MHz carrier frequency modulated with white noise (standard deviation = 0.1, noise bandwidth = 1.0 MHz, modulation index = 0.3) and multipath delay, τ = 13 ns was used. MP – Multipath.

To simplify the control transfer function, H(s), the slopes between S_δ and S_ε can easily be matched by taking the time derivative of $\varepsilon(t)$ to produce $\varepsilon'(t) = d\varepsilon(t)/dt$. This is shown in Fig. 3 (a) indicating that S_δ, $S_{\varepsilon'}$ and $S_{\delta\varepsilon'}$ all now have the same slope. Here, $S_{\varepsilon'}$ is the PSD of $\varepsilon'(t)$ and $S_{\delta\varepsilon'}$ is the CPSD between $\delta(t)$ and $\varepsilon'(t)$. H(s), which maps amplitude fluctuations to multipath error is generated with the LabVIEW frequency response function (FRF). For calculating H(s), $\varepsilon'(t)$ is used as the stimulus signal and $\delta(t)$ as the response signal. Once the transfer function is created, $\varepsilon'(t)$ is filtered with the transfer function and applied to the FM demodulated signal in a feedforward fashion. The PSD of the multipath error $\delta(t)$ is measured with and without the control signal and is shown in Fig. 3 (b). An improvement of greater than 20 dB over two decades of offset frequencies can be seen. However, this scheme works for a fixed multipath condition and requires re-calculation of the transfer function when multipath environment changes (i.e. the antennas positions move). This problem can possibly be addressed with an adaptive control system. Also, in case of two-way ranging conditions a known modulation symbols can be used for H(s) generation.



(a)

(b)

Fig. 3. (a) Plot of the power spectral density $S_{\delta\varepsilon'}$ (cross) along with S_δ (multipath error) and $S_{\varepsilon'}$ (derivative of AM). (b) Plot of S_δ (multipath error) with and without feedforward correction. MP – Multipath.

III. Conclusions

We proposed a new multipath mitigation technique suitable for narrowband systems. It relies on FM-AM correlation when the FM signal experiences multipath, we reported more than 20 dB reduction of multipath distortion.

Acknowledgment

The authors acknowledge Dave Howe and Fabio da Silva for useful discussions and suggestions on this manuscript. This work is sponsored by the Public Safety Communications Research Division at NIST.

References

[1] E. D. Kaplan, Understanding GPS principles and applications. New York: John Wiley and Sons Inc, 2005.

[2] F. Zafari, A. Gkelias, K. K. Leung, "A Survey of Indoor Localization Systems and Technologies," arXiv:1709.01015v2, 14 Mar 2018.

[3] J. Sandber, "Extraction of Multipath Parameters from Swept Measurements on a Line-of-Sight Path," IEEE Trans on Antenna and propagation, vol. AP-28, no. 6, November 1980.

[4] C. Zhang, W. Qi, P. Liu, and L. Wei, "Multipath cancellation by frequency diversity: a training-free and analytical approach to accurate RSS ranging in ground-deployed wireless sensor networks," IET Electron. Lett., vol. 50, no. 6, pp. 471-473, 2014.

[5] A. Hati, C. W. Nelson, and D. A. Howe, "Oscillator PM Noise Reduction from Correlated AM Noise," IEEE T. Ultrason. Ferr., 63, pp. 463-469, 2016.

Best in Class: Leveraging Robot Performance Standards in Academic Competitions to Encourage Development and Dissemination

Raymond K Sheh[1,2] and Adam S Jacoff[2]

ABSTRACT

Standard test methods and academic competitions share much in common. We detail how we use standard test methods to promote education, research, development, and dissemination among the academic community. Since 2014, we have used competitions and open source robot designs to focus students, and particularly high school students, on the challenges of emergency response and public safety robotics. Our two main initiatives are the Rapidly Manufactured Robot Challenge (RMRC), which forms part of the RoboCup Rescue Robot League (RRL), and the Open Academic Robot Kit (OARKit). The RRL and RMRC leverage Standard Test Methods for Response Robots, developed under ASTM International Subcommittee E54.09 on Homeland Security Applications: Response Robots. The standards developed under ASTM E54.09 are significantly supported by the pre-normative research collaboration between the US Department of Homeland Security (DHS) and the US National Institute of Standards and Technology (NIST). These standard test methods are an effective way of communicating the challenges of the domain and focusing research and development on open problems. By measuring the performance of prototypical implementations in a consistent, comparable manner, standard test methods also allow students to compare performance with each other, as well as with commercial, deployed robots. The OARKit aims to capitalize on the ease of comparison and collaboration that comes with the use of standard

[1] Department of Computing, Curtin University, Bentley, WA, 6102, Australia; 0000-0002-9693-3184
[2] Intelligent Systems Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA

Sheh, Raymond; Jacoff, Adam. "Best in Class: Leveraging Robot Performance Standards in Academic Competitions to Encourage Development and Dissemination." Paper presented at ASTM STP1614 on Symposium on Homeland Security and Public Safety: Research, Applications and Standards, San Diego, CA, US. June 28, 2018 - June 29, 2018.

test methods by lowering the resource and expertise barriers for entry into response robotics research. This family of robot designs form ideal starting points for new and existing teams to enter the RMRC. Teams build these robots by following basic instructions and then improve them in their area of expertise. The results are then rigorously measured in the competition and disseminated to other teams. Thus the community of teams share each other's developments and push the state-of-the-art forward.

Keywords

Robot Competitions, Response Robotics, Education

Introduction

Standard test methods share many similar characteristics with competitions, particularly competitions that focus on individuals or teams achieving a score that is then compared with others. In both standard test methods and competitions, performance is measured in a way that is intended to be fair, repeatable, reproducible, and significant in some way.

The RoboCup Rescue Robot League (RRL) has been running since 2000. This research competition, originally for university research students, aims to foster the development of technologies to address gaps in the deployed capabilities of response robots[3], such as those deployed for addressing natural disasters, explosive ordnance disposal, or nuclear incident response. This competition forms an integral part of the development process for the Standard Test Methods for Response Robots, developed under ASTM International Subcommittee E54.09 on Homeland Security Applications: Response Robots. The RRL is an incubator for conceiving,

[3] Certain commercial products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

developing, refining, and validating new test methods, in the presence of prototypical robots that embody new capabilities and ideas. It also disseminates the test methods among the academic and developer communities and provides an opportunity to perform repeatability and reproducibility testing.

The RRL Rapidly Manufactured Robot Challenge (RMRC) is a research competition aimed at high school and early undergraduate students. It also leverages these test methods to expose these younger students to open research challenges in the field of response robotics. Since 2014, we have been developing the RMRC alongside the Open Academic Robot Kit (OARKit), an initiative that aims to bring interesting research-level robots into high school and undergraduate classrooms. It lowers the barrier of entry in terms of cost and resources by leveraging open source principles and low cost, rapid manufacturing and prototyping equipment such as 3D printers and laser cutters. In this paper, we present the history, motivation, and latest developments of these two initiatives and their close relationship to the ASTM E54.09 standard test method development efforts.

Robotics competitions bring significant benefits to younger students and research communities alike. For example, they provide inspiration to students, both by way of the application as well as through observing their peers from the other side of town or the other side of the world. They are a conduit for disseminating information about the challenges and best-in-class solutions to those challenges. They guide the thinking of students towards gaps in current capabilities and provide a way to measure their progress towards addressing them. They also bring to the attention of industry and government the capabilities being developed in classrooms and institutions that may otherwise not be apparent, and present them in a way that is directly comparable to the deployed state-of-the-science.

Standard test methods can play an important role in amplifying these benefits by ensuring that the challenges that guide and inspire the students direct them towards measurable deficiencies in the current state-of-the-art. The Standard Test Methods for Response Robots are designed to be elemental, abstracted tasks that are directly relevant to the real world tasks that response robots must perform in applications such as search and rescue, explosive ordnance disposal, hazardous and nuclear waste cleanup, disaster survey, military reconnaissance, and the like.

They do not replace testing of robots in operational scenarios; rather they provide an indication of the strengths and weaknesses of the robotic system and are a filter to determine which systems are ready for operational testing. To use a sporting analogy, they are like the basic tests of running, jumping, catching, throwing, and so-on that a coach might use to determine how their team is performing and where individual players may need to improve. These tests would also be used by the coach to filter new players before they play their first test game.

THE HISTORY OF ROBOCUP

Since 1997, the RoboCup Federation has been holding robotics competitions. These first focused on the Artificial Intelligence (AI) research community. In the 20 years since, it has branched out across both the wide array of robotics-related topics as well as the various age groups, from junior school students through to early career researchers.

The objective of RoboCup is that “By the middle of the 21st century, a team of fully autonomous humanoid robot soccer players shall win a soccer game, complying with the official rules of Fédération Internationale de Football Association (FIFA, the world soccer governing organization), against the winner of the most recent World Cup.” [1]. Conceived as a “Grand Challenge”, RoboCup is a vehicle to promote robotics and AI research. It is also a type of “standard problem” that can be replicated around the world, with rules that are familiar to many around the world and are well-documented and understood.

The RoboCup Rescue Robot League

In 2000-2001, the RoboCup Rescue Robot League was introduced. This competition, first piloted at the Association for the Advancement of Artificial Intelligence (AAAI) 1999-2000 conferences, brought the challenge of the Urban Search and Rescue task to university students, who were tasked with building a robot that could survey a disaster site, find victims, determine their condition, and build maps of the environment [2]. In these early years, the arena consisted of various random items of furniture, debris, and other household items, arranged into three regions that represented different mobility challenges [3] [4].

The competition was run in three rounds. First, a preliminary round is run with all teams being given multiple opportunities to search and survey the arena. The arena was often split in half, allowing two teams to run at the same time before switching sides. Teams scored points by finding mannequin body parts as simulated "victims", and other objects of interest, overcoming debris and other obstacles strewn around the arena. The points were added up and a clear break in the scores was sought that separated out the top 5-10 teams, thus avoiding the situation where a team "just" missed out. These teams progressed to the finals. Teams were given the whole arena to search and survey, with the team that found the most victims and brought back the most information was declared the winner.

During these two rounds, teams were prevented from observing the arena until after their runs, thus the competition was truly a search task. While realistic, this competition was less reliable as a standard test due to the highly random nature of the behavior of the robots, as well as the element of luck involved in searching the arena.

A third round, called Best-in-Class, allowed teams that specialized in particular aspects of the competition to demonstrate their capabilities, even if they did not have the broad base of expertise that would allow them to win the overall competition. Initially, Best-in-Class awards were given for Mobility and Autonomy. These focused on the robot's ability to traverse terrain and to navigate autonomously, respectively. In later years, this was joined by Manipulation, which focused on robots with the ability to manipulate objects.

Starting in 2005, in conjunction with the launching of the ASTM standardization effort supported by DHS, prototypical standard test method elements began being introduced into the arena. This started with the Random Stepfields [5] shown in Figure 1, a repeatable, reproducible analog for unstructured terrain. The random mannequin body parts that were the simulated "victims" in previous years were also replaced by standard artefacts, designed to test the abilities of the robots' various sensors, such as thermal and visible light cameras, microphones, carbon dioxide sensors, and so-on. In subsequent years, the arena shifted to being based on standard test method apparatuses, primarily for terrain traversal such as crossing ramps, symmetric stepfields, and stairs. Thus, as teams ran their robots through the arena, they were also, in effect, performing informal tests within the standard test methods.

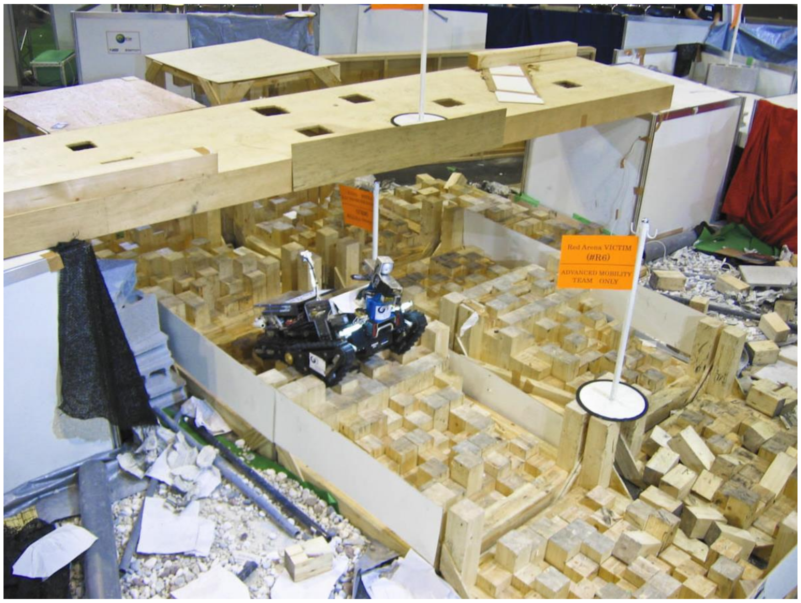


Figure 1: The Random Stepfields at the RoboCup Rescue Robot League 2005.

To further reduce the element of chance in the competition, in the late 2000's the secrecy component of the competition was also dropped, with team members being allowed to enter the arena prior to the run and verify that it had been re-set correctly after the previous team's run. This had the effect of switching the competition from a search task to one of focusing on the test methods integrated into the arena. The more test methods teams could overcome, the more "victim" sensor test boards they could reach and thus more points they could obtain. At the same time, the Best-in-Class competitions were introduced as separate rounds within the standard test method apparatuses. Most obvious of these was the Best-in-Class Mobility competition, which was a direct application of ASTM E2828-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Symmetric Stepfields [6].

Around 2015, the competition progressed further towards standardization. Instead of running one or a small number of standard test method rounds as a Best-in-Class round, the preliminary rounds were replaced by standard test method runs within separate lanes. The goal became that of statistically significant testing in the preliminary rounds with as many teams as possible running tests in parallel. The scores for each test were normalized according to the best performing team in that test and then combined to determine which teams progressed to the final rounds.

Outreach Activities

The RoboCup Rescue Robot League goes beyond just a competition. The overall goal of the League is to advance the state-of-the-science in response robotics, in part by fostering the sharing of capabilities between teams. The competition itself goes some way towards this by allowing teams to observe each other's capabilities, as tested in comparable test methods. However, it can be quite difficult for teams to share information during competition, if only because time and development effort pressures mean that many teams, especially those who are performing particularly well, would rather focus on their next competition run.

As a result, the RoboCup Rescue Robot League also hosts regular teaching camps [7] or summer schools. The League community is also active in academic conferences, particularly the RoboCup Symposium and the International Symposium on Safety, Security and Rescue Robotics (SSRR), run by the Institute of Electrical and Electronic Engineers (IEEE) Robotics and Automation Society (RAS). These venues provide a more relaxed environment in which those who developed the Best-in-Class capabilities of the previous competition can present their work.

The teaching camps in particular are structured to be highly practical. Usually half of the time is set aside for traditional lectures and presentations, while half of the time is spent in multi-tracked practical sessions where smaller groups are led in the implementation, or re-implementation, of a capability that they can then take and build upon or use as a point of comparison. This provides opportunities for other teams to bring their own robots and learn, in a hands-on fashion, how to implement these capabilities. Examples of capabilities that have been spread through the league through these events include various 2D and 3D Simultaneous Localization and Mapping (SLAM) implementations [8], inverse kinematics, aerial robots, and user interfaces.

These events also form a valuable opportunity to form collaborations with stakeholders outside of traditional university academia. The teaching camps and summer schools are usually hosted at a site where responders train and always have a contingent of responders in attendance, such as firefighters and police officers. These personnel attend alongside the students, with the goal of sharing their knowledge of the application, grounding the students' understanding of where their technologies may be useful, as well as to learn about new capabilities that may be of use to them but may not be well known outside the laboratory. Other academic sectors have also attended these events to learn about potential opportunities to expand their activities. Indeed, the RMRC was launched at the 2014 Safety, Security and Rescue Robotics Summer School and Workshop [9], which was jointly hosted by the Perth Artifactory makerspace and Curtin University. For the first time, a contingent of high school teachers joined the event and formed the core teams that participated in the pilot competitions in the following years.

Finally, these events provide an opportunity to analyze the outcomes of the prototypical test methods, experiment with them in a more informal setting, and document them in preparation for potential standardization. They also provide an avenue to disseminate in more detail the test methods among the broader academic community.

COMPETITION TRADE-OFFS

In discussing robot and AI competitions and their ability to advance research goals, we find it useful to consider competitions according to two different trade-offs. The first is the task, be it abstract or real-world application based. The second is expected novelty, be it known or research.

Competition Task Trade-Off

An example of an abstract task is robotic chess. This is a game, developed for a particular purpose. On the one hand, while the research challenges are still very real and applicable, the game is heavily abstracted from open real-world challenges. On the other hand, the problem is also very well defined with a well understood set of semantics, problem definition, and evaluation metric.

In contrast, a real-world application based task might be the US Defense Advanced Research Projects Agency (DARPA) Urban Challenge [10] which involves autonomous robotic vehicles driving through an urban area, complete with other autonomous and human-operated vehicles. The top performers in such a competition have produced a working solution that solves an open real-world application problem. Of course there is usually still significant work to be done to make their entry viable in the real world from safety, economic, production, and regulatory perspectives (among others). However, administering and competing in such competitions tend to be very resource intensive. This limits both the ability to run such competitions as well as the ability of teams to compete.

All competitions trade off between these two characteristics depending on their goals, a compromise that is generally mutually exclusive. Our goal with the RRL and RMRC is to balance being abstract enough to make it easy to administer and compete in these competitions, while also being close enough to the application that they can foster and evaluate the development of technologies that are applicable to real-world problems in the short to medium term.

We leverage the Standard Test Methods for Response Robots as a way to achieve a balance between these two characteristics. Test methods, by their nature, must be both abstract enough to be reproducible and repeatable, scientifically rigorous, and economically viable to reproduce and disseminate, and yet the resulting metrics must clearly be an accurate reflection of a real-world need. By building a competition based around standard test methods, we take advantage of all of the development effort, industry and end-user consultations, and international experience behind these test methods; resources that are generally unavailable to most competitions.

Competition Novelty Trade-Off

Competitions also vary in terms of the novelty of the challenges and expected solutions. On the one hand are competitions where the underlying problems are relatively well understood. Teams are differentiated by how well they can answer this problem, which may still require some degree of novelty, but in general what is required of teams is well understood. Examples include track racing or flat-floor mazes. Such competitions encourage refinement and optimization of solutions.

On the other hand are competitions that pose problems for which there is no known satisfactory solution. Such competitions encourage a much wider variation in solutions, but these solutions also tend to be less well optimized and more experimental. The RoboCup Rescue Robot League is an example of an application that tends towards this end of the spectrum.

RELATED WORK

Research competitions and open source hardware initiatives have been gaining significant traction over the last decade as their value to both teaching and research have become recognized. In this section we will discuss some related competitions and initiatives.

Related Competitions

Robotics competitions have been a feature of high schools for many decades as teachers have leveraged their ability to consolidate the many aspects of what is now called STEAM – Science, Technology, Engineering, Art, and Mathematics – into an event that inspires creativity and problem solving. As a result there is currently a plethora of different competitions, many of which are surveyed in [11].

RoboCup itself is a family of competitions for participants from 14 years old (some regional competitions have lower limits) up to early career researchers (and beyond). In the Junior category, which allows students up to 19 years old, there are three main competitions: OnStage (performance), Soccer, and Rescue. There are also additional challenges such as the CoSpace challenge [12] as well as the RMRC itself. Apart from the RMRC, these competitions are based on games and abstract tasks.

Beyond RoboCup, perhaps one of the best known families of robotics competitions in the present day are the "For Inspiration and Recognition of Science and Technology", or FIRST, Robotics Competitions [13]. Catering to students from kindergarten through high school, the various competitions that make up FIRST aim to not only inspire students to partake in the study of science and technology, but to also teach students about other life skills such as "self-confidence, communication, and leadership".

To this end, RoboCup Junior and FIRST quite deliberately pick somewhat abstract games, made up of combinations of tasks such as moving balls into goals, navigating mazes, pushing levers, and delivering objects on a flat playing field. The point of the competition isn't the solution to the problem itself, but rather the journey that students take in solving the problem. Competitions such as RoboCup Junior OnStage also focus on artistic aspects, particularly as it relates to performance art. CoSpace augments the real robotics competition with a parallel virtual robotics environment to enable more advanced programming and algorithm development. FIRST augments these problem solving tasks with those that encourage teamwork, professionalism, and community service, such as having to form ad-hoc teams during the competition and giving presentations. Indeed, having a more concrete, real-world problem could arguably get in the way of these goals.

Traditionally, competitions where the goal is the solution, rather than the journey, has been the realm of more senior students. These competitions pose open research problems and teams that do well will have developed a novel contribution to a real-world problem. Examples of such competitions include the DARPA Grand and Robotics Challenges, the World Robot Summit 2020 [14], and the RoboCup Rescue Robot League [15]. These competitions tend to require more resources to administer and compete in, both in terms of materials as well as expertise.

In a sense, developing these competitions is very much like developing standard test methods. The goal is to strike a balance in developing a test that is sufficiently abstract that it can be reproduced, understood, and disseminated reliably and fairly, while also being relevant to the real world task, with all the variability that this entails.

Related Open Source Robotics Initiatives

One way to lower the material and expertise cost of participation for new and existing teams into any of these competitions is to allow the teams to build on others' work. Different initiatives have been proposed to assist in this task.

Some competitions make use of specified or mostly specified parts. Examples include competitions that make use of construction kits, such as the FIRST Lego League and several of the RoboCup Junior competitions [16]. Other competitions, such as the FIRST Robotics Challenge [13], provide a specific, common kit of parts along with documentation and examples.

Construction kits that combine physical components with sensors, actuators, and computation, such as Lego Mindstorms, have been pivotal in bridging the gap between computer science and physical robots in the high school because they remove a lot of the complexity involved in actually building the robots that are necessary to embody code and algorithms. The use of such kits is also significantly safer for younger students than building robots in a more traditional manner out of metal, wood, or plastic. However, it is limited in its ability to build larger, more complex, more durable robots due to the need to limit designs and structures to those that are compatible with the particular build system. It is also difficult to teach more advanced concepts in design, particularly as it relates to design for manufacture. Dissemination and re-use of other teams' work can also be challenging with these kits due to the high level of design coupling between different parts of the robot. Thus, while such kits are a tremendously valuable tool to teach introductory robotics, it is difficult to build larger, more complex robots and to leverage other students' work.

In recent years, 3D printing and laser cutting have become much more widespread among high schools, particularly as open source equipment and software have lowered the financial and expertise cost. Practical, low cost 3D printers capable of printing robot components out-of-the-box, with minimal expertise, can be purchased for less than $300 USD, while larger printers, capable of printing whole robot parts without needing to glue them together afterwards, can be purchased for less than $1,000 USD. In a classroom environment, laser cutters have also become more commonplace and can often be more practical as they can produce larger (albeit flat) components much more quickly than 3D printers. Here too, prices have fallen with easy to use units of a practical size starting from $2,500 USD.

Recent years have also seen rapid advancements in the power, flexibility, availability, and ease of use of open source electronics such as Arduino [17] and Raspberry Pi [18] development boards. These advancements are in large part due to the fact that, with open source projects such as these, anyone can make improvements based on their area of expertise and share them with the broader community. As a result, it has become even easier to build embedded computation, advanced sensing, and reliable communications into classroom projects.

The widespread availability of these systems has resulted in a plethora of open source robotics initiatives, including our Open Academic Robot Kit [19]. The goal is to enable ad-hoc collaboration between students in different parts of the world and across year levels. For example, we have developed a basic platform that is intended to be easy to build upon, complete with basic electronics, sensors, and code [20]. A PhD student from Australia might implement a new vision algorithm that allows the robot to avoid obstacles using its camera. A high school student in Thailand might add a better arm design, while a makerspace group in the United States might contribute better wheels. These groups – and others – can benefit from each other's work.

Other open source robotics designs of interest to high school students include the Robotis TurtleBot 3 [21], Poppy [22], Vorpal Robots [23], and Yale OpenHand [24]. All of these designs share these common characteristics of providing students with a basic design with all components either 3D printed or easily available off-the-shelf, plus instructions and source code that allows them to "close the loop" on building an interesting robot that they can then extend. Some of them even have "standard" tasks that can be used as the basis of student competitions.

The Rapidly Manufactured Robot Challenge

The RMRC, as a sub-competition within the RRL, is designed to foster research and development among young students, including contributions to the open source community. The design of the competition and arena is therefore focused on providing students with opportunities to experiment, specialize in particular areas of interest, demonstrate their capabilities, and share these capabilities back to the competition and research community.

COMPETITION STAGES

The RMRC starts with teams qualifying based on their Team Description Materials (TDMs). Teams that are selected based on these materials come to the competition and proceed through two rounds. During the initial, preliminary round, the Standard Test Methods for Response Robots are laid out individually. Teams select test methods in which to evaluate their robots with the aim of achieving the highest metric possible within a prescribed 10 minute period. This is followed by one or more final rounds where the test methods are arranged in a maze or course with teams scoring points according to how many test methods they overcome in sequence.

Qualification

Teams are qualified to compete in the World Championships of the RMRC via three avenues. First, teams may qualify by submitting TDMs. These correspond to the Team Description Papers that are requested of the Major RoboCup Rescue Robot League teams (who are mostly undergraduate or graduate students), but in a more general form that makes it conducive to integration into high school media curriculum such as a website, blog post series, and/or videos. Teams may also be selected if they reached the finals in the previous year or if they place in regional competitions.

The TDM is required to cover the following topics, regardless of the format in which it is presented.

- Logistical info: Team Name, Organization, Country, Contact details, Website.

- Introduction summarizing:
    - The team.
    - The technical aspects that it focuses on.
- System description, describing:
    - Hardware.
    - Software.
    - Communications.
    - Human-robot interface.
- Application, describing:
    - Setup and packup of the robot and operator station.
    - Mission strategy.
    - Experiments and testing that have been done or will be done.
    - How the particular strengths of the team are relevant to applications in the field.
- Conclusion, summarizing:
    - What the team has learned so far.
    - What the team plans on doing between now and the competition.
- Appendix containing:
    - One table per robot outlining the components and estimated cost of the robot.
    - At least one picture, 3D rendering, or technical drawing of the robot.
    - A list of software packages, hardware, and electronic components that have been used, or plan to use, particularly those from the Open Source community, through the Open Academic Robot Kit or otherwise.
    - A list of software packages, hardware, and electronic components and designs that have

been, or plan to be, contributed to the Open Source community, through the Open Academic Robot Kit or otherwise.

- References (to other work that the team has made use of).

These materials are evaluated by a panel of high school and university academics. Teams are qualified not just for the potential to perform, but also for the potential to learn and for excellence in specific capabilities.

Preliminary Round

Teams that come to the World Championship first enter a preliminary round that runs over two full days. The goal of this round is to give teams the most number of opportunities to demonstrate their capabilities on the test methods that they are able to complete. This includes giving them time to experiment, to fail if necessary, and re-attempt.

The test methods are divided into two categories: those that relate to the ability of the robot to reach its destination and those that relate to its ability to perform a task once at the destination. We will refer to the former as the traversal test methods and the latter as the sensing and manipulation test methods, both of which are critical in measuring the performance of these systems as a whole, as they relate to their ability to perform in the response robotics application. The total scores achieved in each of these categories are multiplied together to yield the teams' final score.

Traversal test methods

The traversal test methods represent the ability of the robots to reach their destination in order to perform a task. These are generally maneuvering or terrain tasks such as centering between obstacles, climbing stairs (ASTM E2804-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Obstacles: Stairs/Landings [25]), or overcoming rough terrain (e.g., ASTM E2826-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Continuous Pitch/Roll Ramps [26], ASTM E2827-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Crossing Pitch/Roll Ramps [27], and ASTM E2828-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Symmetric Stepfields [6]). These are all presented in the form of terrain pallets as shown in Figure 2. Each has a start and an end with the traversal challenge in the middle and points are scored for reaching each end of the pallet. One exception is the nodal manipulation (“Pipestar”) apparatus, which also exists in the arena as a specific test for manipulation-focused teams as shown in Figure 3. Robots may perform one or more of five tasks at each pipe for one point each: Precision touch (ASTM WK54272 Evaluating Ground Response Robot Dexterity: Touch or Aim), Rotation of the object (ASTM WK54273 Evaluating Ground Response Robot Dexterity: Rotate), Extraction of the object, Insertion of the object (ASTM WK54274 Evaluating Ground Response Robot Dexterity: Extract and Place), and Inspection of the interior of the pipe (ASTM WK54271 Evaluating Ground Response Robot Dexterity: Inspect). Approved standards are indicated by a label that begins with an "E" whereas draft test methods are considered work items and indicated by a "WK". These draft test methods are being refined in preparation for balloting as standards [28].

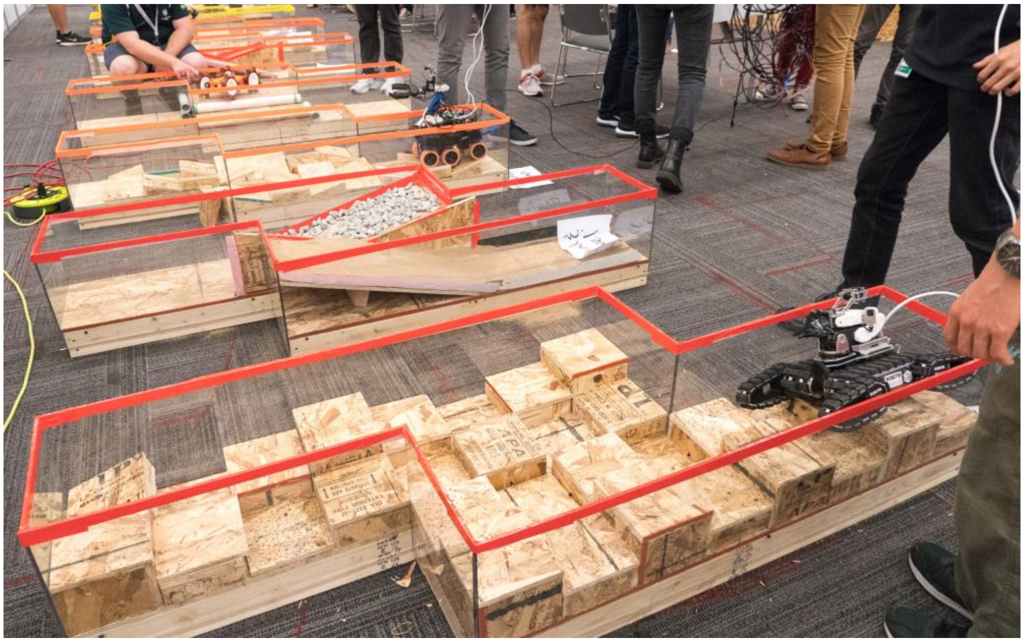


Figure 2: Terrain test method apparatuses arranged as individual lanes for preliminary trials.

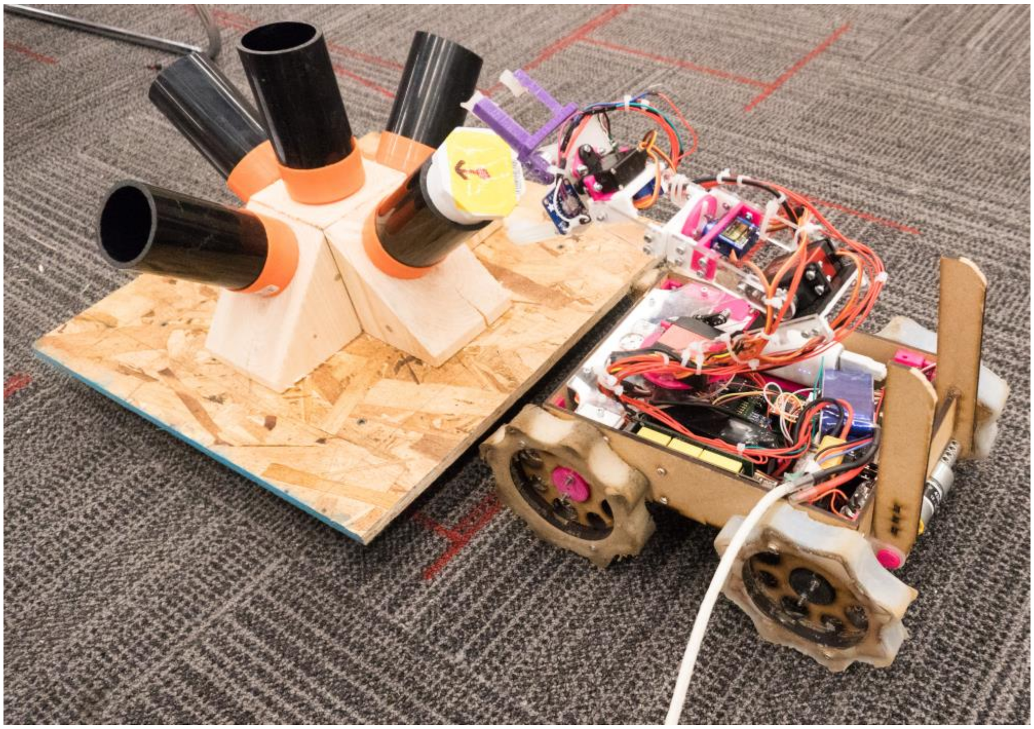


Figure 3: The Nodal Manipulation ("Pipestar") apparatus.

The tests are performed according to the ASTM International standard test method procedure [29] with the metric being the number of half-laps of the terrain pallet that the robot performs within the prescribed time, usually 10 minutes. These pallets are arranged individually as shown in Figure 2 so that they can be run in parallel. Points are assigned for each test method that a team successfully completes a measurement run in. Teams have the opportunity to try each traversal test method several times with their maximum metric for each test method counting towards their final total. Each test method's scores are normalized such that the top metric in that test method is equal to 100. Thus if a team is the only one to successfully achieve a measurement in a given test method, perhaps because it is particularly specialized or difficult, that team automatically gains 100 points for that test method. In contrast, a team must achieve excellence in an easier or more popular test to gain the same number of points.

A scheduling matrix, an example of which is shown in Figure 4, is drawn up. Teams are invited, in random order, to claim a slot representing a particular test method, at a particular time (in the figure, teams are denoted by a two-letter code appearing in the upper left corner). Once each team has claimed a slot, the cycle repeats with their second slot and so-on. The placement is constrained by the inability of teams to place two of their own slots on the same line (as they can only do one test at once), some test methods cannot run in parallel (because they share equipment), and the total number of taken slots on each line cannot exceed the number of referees available to administer each test. The order is then inverted for the following half-day. Once each team has claimed all the slots that they wish to, remaining spots may be used by any team to practice. Teams may also move slot claims to empty (future) slots or negotiate with other teams to swap. As tests are performed, the results are recorded in each slot, along with the name of the administrating referee.

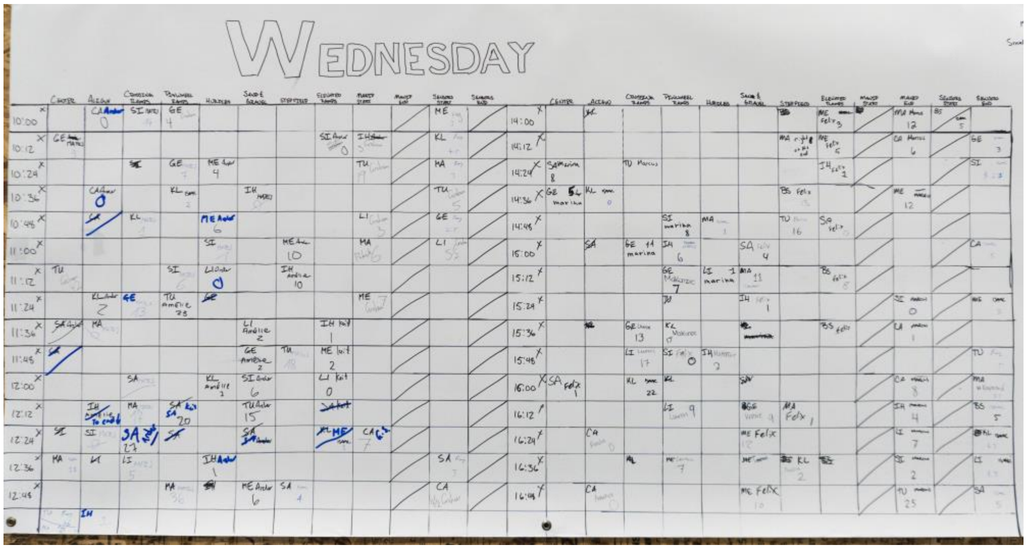


Figure 4: A scheduling matrix for the preliminary (standard testing) rounds.

To the maximum extent possible, teams are provided with the ability to run as many times as their own scheduling and reliability allows. This encourages teams to come prepared and to develop reliable systems. Runs are scheduled every half-day during team leader meetings in the morning and at lunch time. Each run is 10 minutes long with 5 minutes for changeover, yielding four runs per hour. Each half-day runs for around 3-4 hours and there are generally 4-5 referees to administer tests; thus there are generally around 96-160 test opportunities per day or 192-320 test opportunities across the qualification round. In the 2017 World Championships, being the first year that the RMRC was run as an open competition, six viable teams competed yielding over 100 standardized test results. With around 10-13 teams as anticipated going forward, this allows each team between 15 and 30 testing opportunities across the qualification round. There are generally around 10 test methods available, thus teams will generally not be schedule-limited in their ability to try test methods multiple times to achieve their best score.

Sensing and Manipulation test methods

The sensing and manipulation test methods represent the ability of the robot to perform its tasks once it arrives at its destination. These include tests of vision, directed perception, audio acuity, retrieving objects, pushing buttons, and turning valves. These tests are repeated every half-day, during one of the time slots selected in the matrix, to ensure that the capabilities are maintained through the competition. The result is used as a multiplier and applied to the scores of the team's runs for that half-day. This means that teams need to exercise care when driving their robots so that delicate sensors and manipulators (such as robot arms) are not damaged.

Finals

The scores from the Preliminary round are tabulated and around five teams selected. The specific cutoff score is decided by looking for a distinct gap in the preliminary scores. It should be undeniable that the worst performing team that goes through to the finals has performed significantly better than the best team that did not make the cut.

For the final rounds, the traversal test methods are rearranged into a maze as shown in Figure 5, with the arrangement decided in consultation with the teams that qualified. Teams each have between 10 and 15 minutes per run with 5 minute change-overs, depending on how many teams are admitted to the finals. Following a half-day break from the Preliminary rounds for arena reconfiguration and team practice, the Final rounds are run across two half-days. The aim is to allow each team to run at least five times through the arena over the course of the two half-days.

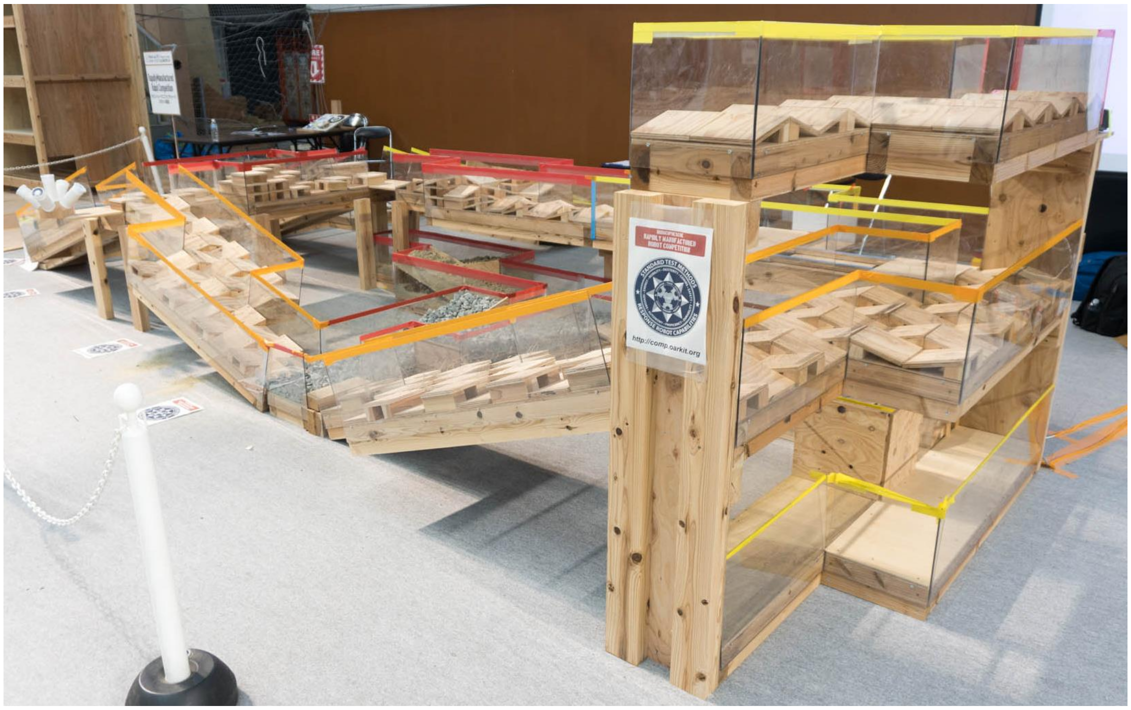


Figure 5: The terrain pallets arranged into a maze for finals.

During each finals run, teams are invited to start their robot anywhere in the arena and the goal is to accumulate as many points through the arena maze as possible. Each terrain pallet yields up to two points, one in each direction, with each point awarded when the robot either touches the end of the pallet or exits onto a connected pallet. As a result, teams need to be strategic in choosing their starting point and their path so as to maximize the number of points while minimizing excessive terrain traverses or the need for excessive risk early in their run. Teams are allowed to drop their worst-performing score, allowing teams to experiment with the different format. This score is then multiplied by their score in the Sensing and Manipulation test methods, which are performed once during the day.

ARENA AND TEST METHODS

The arena is made up of standard test method pallets, arranged first as individual test lanes for the preliminary rounds and then as a maze for the finals. It is important to note that these test method pallets are not simplified versions of the test methods. They are full test methods, built at the 30 cm scale to represent smaller, confined environments such as collapsed buildings, air conditioning ducts, and industrial plants. Currently the following test methods are implemented. Examples of these are shown in Figure 6, in order:

- Center Between Obstacles (no ASTM number assigned yet).
- ASTM WK53649 Evaluating Ground Response Robot Maneuvering: Align Edges.
- ASTM E2827-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Crossing Pitch/Roll Ramps [27].
- ASTM E2802-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Obstacles: Hurdles [30].
- ASTM E2992/E2992M-17 Standard Test Method for Evaluating Response Robot Mobility: Traverse Sand Terrain [31].
- ASTM E2991/E2991M-17 Standard Test Method for Evaluating Response Robot Mobility: Traverse Gravel Terrain [32].
- ASTM E2828-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Symmetric Stepfields [6].
- Elevated ramps (no ASTM number assigned yet).
- Manipulation pipestar (supporting ASTM WK54272, WK54273, WK54274, and WK54271 as described previously).

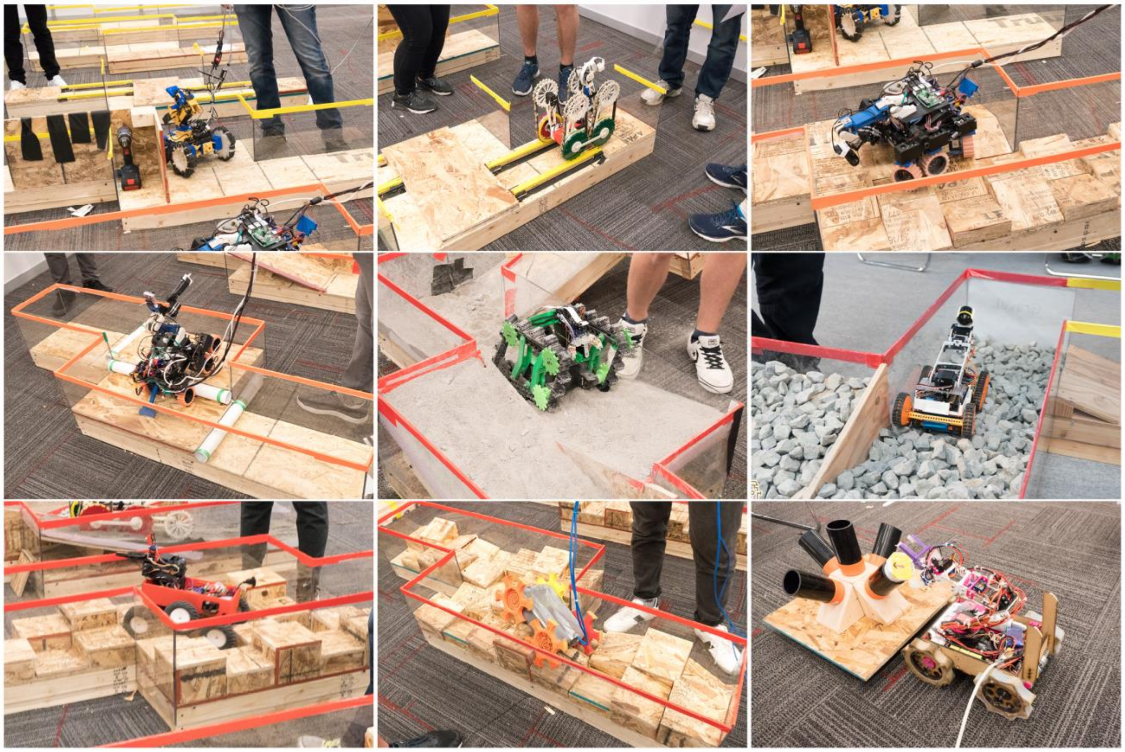


Figure 6: Examples of the different terrains represented in the RMRC, along with examples of rapidly manufactured (3D printed or lasercut) robots.

Sensor tests were embodied by the “Readiness Board”, as shown in Figure 7. This consisted of embedded versions of test methods for ASTM E2566-17a Standard Test Method for Evaluating Response Robot Sensing: Visual Acuity [33], ASTM WK57967 Evaluating Ground Response Robot Sensing: Thermal Image Acuity, ASTM WK60783 Evaluating Ground Response Robot Sensing: Audio Speech Intelligibility, and prototypical test methods for Survey Acuity, Motion Detection, CO_2 detection, and Hazardous Material Label Recognition. These are performed at a standard near-field distance of 40 cm (16 in) and each test is thresholded to represent a point each.

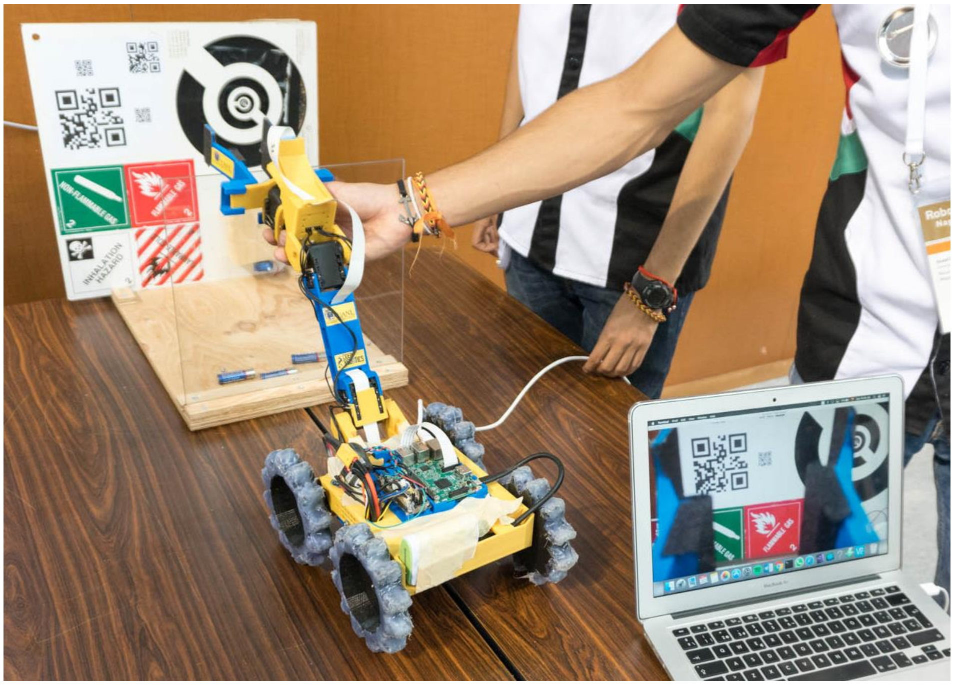


Figure 7: A robot performing the sensor "readiness" test.

Taken together, the terrains, manipulation test, and readiness board represent challenges that robots face getting to the site where they must work, manipulating what they must work on, and then observing it. Thus they each contribute to the scoring. During the preliminaries, there are 12 tests offered.

Four represent mobility and maneuvering challenges that all robots can attempt:

- Center Between Obstacles (no ASTM number assigned yet).
- ASTM WK53649 Evaluating Ground Response Robot Maneuvering: Align Edges.
- ASTM E2827-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Crossing Pitch/Roll Ramps [27].
- A variant on the above called Pinwheel Ramps.

Four represent advanced mobility tasks:

- ASTM E2802-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Obstacles: Hurdles [30].
- ASTM E2992/E2992M-17 Standard Test Method for Evaluating Response Robot Mobility: Traverse Sand Terrain [31], and ASTM E2991/E2991M-17 Standard Test Method for Evaluating Response Robot Mobility: Traverse Gravel Terrain [32]. (combined for the preliminaries),
- ASTM E2828-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Symmetric Stepfields [6].
- Elevated Ramps (no ASTM number assigned yet).

Finally, two represent “payload” tasks:

- Manipulation pipestar (supporting ASTM WK54272, WK54273, WK54274, and WK54271 as described previously).
- Readiness board (supporting embedded variants of ASTM E2566-17a Standard Test Method for Evaluating Response Robot Sensing: Visual Acuity [33], ASTM WK60783 Evaluating Ground Response Robot Sensing: Audio Speech Intelligibility, ASTM WK57967 Evaluating Ground Response Robot Sensing: Thermal Image Acuity, and ASTM WK54755 Evaluating Ground Response Robot Sensing: Match Colors).

The latter two were repeated twice, once at the start of the preliminary day to evaluate the capabilities that robots started with and again at the end of the day to determine if any capabilities had been lost due to damage/wear during the rest of the testing.

The final runs are a combination of tests that together represent a mission. To ease the logistics of the competition, the readiness board is performed at the start and end of the day, representing the capabilities of the robot at the start and after any wear or damage during the day. This score is then multiplied by the score that the robot achieves during its run through the maze, including both the various terrains as well as the manipulation task.

The Open Academic Robot Kit

The aim of the Open Academic Robot Kit (OARKit) is to make it as easy as possible for a new team to enter the RMRC. The initial OARKit designs have all mechanical parts 3D printable on a low cost printer. All other parts are drawn from a relatively small set of components that are easily available by mail order and all designs, instructions, and source code are available online under an open source license. This way anyone, anywhere in the world, can follow the downloadable instructions and build a robot that can be used as a starting point for their competition entry.

The use of 3D printing for the mechanical structure, rather than a standard kit of parts, allows for much greater flexibility in the robot designs and the opportunity for students to learn proper structural design under manufacturing constraints. It also allows designs to be shared more easily as teams seeking to replicate a design need only own a suitable 3D printer and a cache of standard parts.

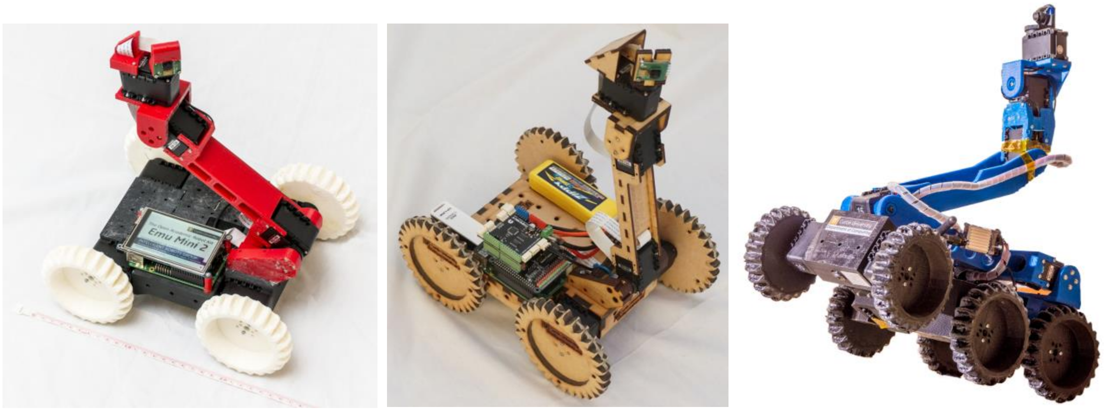


Figure 8: The initial robot designs from the Open Academic Robot Kit.

The first two designs in the OARKit are the “Emu Mini 2” (3D printed and lasercut) and the “Excessively Complex 6-Wheeled Robot”, both shown in Figure 8. These robots are not intended to be the best, or even particularly good, robots for tackling all of the challenges in the RMRC. Rather, they are intended to be good for building additional capabilities from. Over the four years since the OARKit project started, many teams have built variants. Most of these variants share the same parts as the original OARKit: Raspberry Pi for the computation, Arduino for embedded interfacing, and Dynamixel AX-12 servos for motion.

First published in 2014, the OARKit has already spawned a plethora of successors across high schools in Australia, the US, and Europe. Examples of these have appeared at RMRC competitions in the intervening years and some are shown in Figure 9.

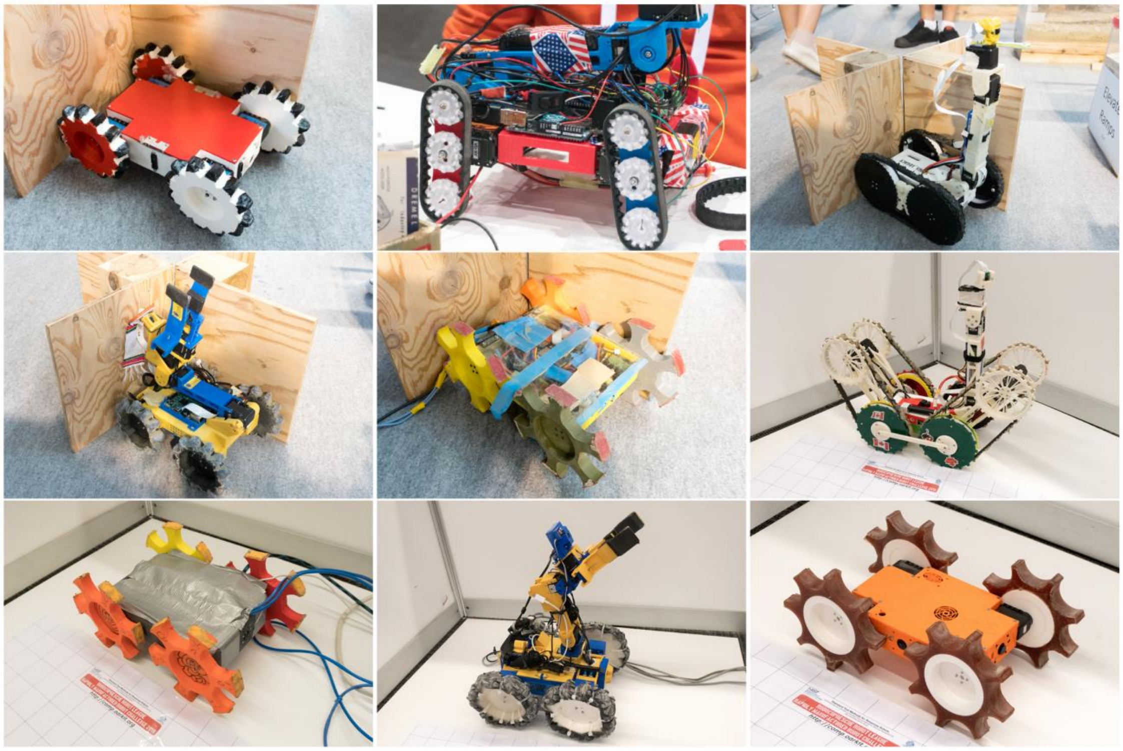


Figure 9: Examples of robots that were based (at least in part) on the original Open Academic Robot Kit designs.

Results

The RMRC has been run as an open competition during both the 2017 and 2018 RoboCup Rescue Robot League World Championships, held in Nagoya, Japan and Montreal, Canada, respectively. For the 2017 competition, 13 teams applied of which 10 were qualified and participated in the competition. Preliminary rounds were conducted over one and a half days, resulting in a total of 139 standard tests and five teams qualifying for the finals.

In 2018, 19 teams applied of which 13 were qualified either on the basis of team description materials or performance in the 2017 competition. 11 of those participated in the competition, completing a total of 297 standard tests over two and a half days and again qualifying five teams for the finals. The results of the 2018 preliminary competition are shown in Figure 10. To provide some anonymization, robots have been denoted by two-letter codes. Each robot's performance in each test was normalized such that 100 represented the (possibly equal) best performance in that test. The radar charts are ordered in descending overall performance and show how the performance of the robots change, both in terms of overall performance as well as in specific areas.

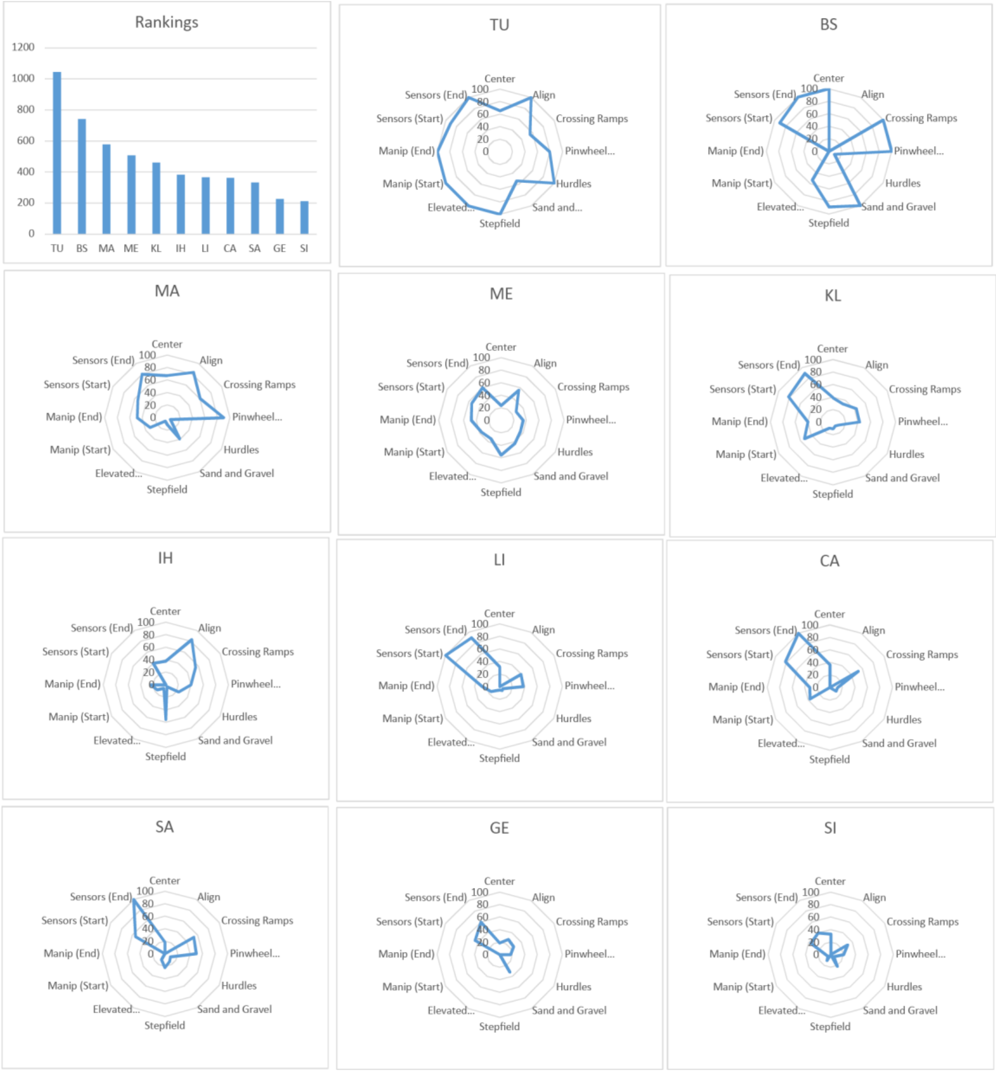


Figure 10: Results of each robot after the preliminary rounds of the 2018 competition. Radar plots for each test are normalized such that 100 represents the best performance in that test. Robots anonymized to two-letter codes.

CONCLUSIONS AND FUTURE WORK

The RMRC is still very much in development as a competition, after two years of open competitions and a preceding two years of trials. Yet we have already observed tremendous growth and maturity in the international community of teams that have developed around it over the last few years. In 2019, we will once again hold the RMRC, in conjunction with the broader RRL, in Sydney, Australia. There we will be refining the rules of the competition as well as the procedures of the test methods, in collaboration with our ASTM E54.09 colleagues. In particular, we will be addressing some issues relating to robot size, whereby several of the robots in the competition barely fit in the mobility test methods and thus the walls became a significant influence on their behavior. This would suggest that these effects either need to be included in the robot performance or otherwise some accommodation made to ensure that their influence is minimized.

Beyond the test methods, we will be providing opportunities for the RMRC and RRL to become more closely integrated again. When the confined space arena, the precursor to the RMRC arena, was first conceived, it was intended to be a “shortcut” into the more difficult areas of the RRL arena for smaller robots. This represented the access options that smaller robots might have in a real response situation that larger robots may not be able to take advantage of. Our plan in 2019 will be to introduce an intermediate arena of test methods at the 60 cm (24 in) scale that will be connected to both the RMRC and RRL arenas for the finals. This will allow robots from both competitions to attempt an intermediate scale before possibly transitioning into the others’ arena. The introduction of the intermediate sized arena may also address some concerns about the robots being too big for the arena, albeit with the disadvantage of making the arena larger and thus less cost effective for high school teams to reproduce.

Over the past several years, we have seen the original Open Academic Robot Kit robots proliferate and further develop as teams improve and share the designs. To further encourage contributions to the open source community, we will be adding a multiplier onto teams' preliminary scores, based on a report, document, or other resource that teams submit two weeks prior to the competition. This resource will be scored based on how useful it is for another team to replicate a particular feature or innovation that the team has developed.

The introduction of standard test methods into the RRL and associated RMRC has accelerated the development of robotics for emergency response and public safety. They have helped to communicate the challenges of the application to students and researchers. In turn, they have helped to communicate the prototypical capabilities within academia to the broader user and manufacturer community. In the process they have been critical in forming an interconnected community of researchers, manufacturers, users, and test developers. Further, the competitions have turned out to be ideal proving grounds for the development of standard test methods which has enabled the rapid passage of relevant standards through ASTM E54.09. This symbiotic relationship is a potential model other standard development efforts could use to educate, speed research and development, and disseminate standards through their communities.

ACKNOWLEDGMENTS

We would like to acknowledge the support of all of the mentors and students who participated in, supported, and contributed to this competition and initiative, without whom it would not be possible. We would also like to thank the volunteer referees and administrators who helped us to run the events and all of our collaborators under the ASTM International Subcommittee E54.09 on Response Robots (and the precursor Subcommittee E54.08.01) for their efforts in developing the underlying standard test methods. Finally we would like to thank our colleagues, particularly

William Billotte, Kevin Jurrens, Kenneth Kimble, and Elena Messina for their assistance in reviewing this paper. This work was performed in part under the financial assistance award 70NANB17H265 from the U.S. Department of Commerce, National Institute of Standards and Technology.

References

[1] RoboCup Federation, "The Objective of RoboCup," 27 June 2018. [Online]. Available: http://www.robocup.org/objective.

[2] R. Sheh, S. Schwertfeger and A. Visser, "16 Years of RoboCup Rescue," *KI - Künstliche Intelligenz,* vol. 30, no. 3-4, pp. 267-277, 2016.

[3] A. Jacoff, E. Messina and J. Evans, "Jacoff, Adam, Elena Messina, and John Evans. "Performance evaluation of autonomous mobile robots," *Industrial Robot: An International Journal,* vol. 29, no. 3, pp. 259-267, 2002.

[4] A. Jacoff, E. Messina and J. Evans, "A standard test course for urban search and rescue robots," NIST Special Publication SP, Gaithersburg, 2001.

[5] A. Jacoff, A. Downs, A.-M. Virts and E. Messina, "Stepfield pallets: repeatable terrain for evaluating robot mobility," in *PerMIS '08 Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, Gaithersburg, Maryland, USA, 2008.

[6] *ASTM Standard E2828-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Symmetric Stepfields*, West Conshohoken, PA: ASTM International, www.astm.org, 2011.

[7] R. Sheh, B. Collidge, M. Lazarescu, H. Komsuoglu and A. Jacoff, "The response robotics summer school 2013: bringing responders and researchers together to advance response robotics," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, 2014.

[8] S. Kohlbrecher, K. Petersen, G. Steinbauer, J. Maurer, P. Lepej, S. Uran, R. Ventura, C. Dornhege, A. Hertle, R. Sheh and J. Pellenz, "Community-driven development of standard software modules for search and rescue robots," in *Proceedings of the 10th IEEE International Symposium on Safety Security and Rescue Robotics (SSRR 2012)*, 2012.

[9] R. Sheh, "RAS Summer School Held in Australia [Society News]," *IEEE Robotics & Automation Magazine,* pp. 134-136, 3 2015.

[10] M. Buehler, K. Iagnemma and S. Singh, The DARPA Urban Challenge: Autonomous Vehicles in City Traffic, Springer, 2009.

[11] A. Eguchi, "Bringing Robotics in Classrooms," in *Robotics in STEM Education,* Springer, Cham, 2017, pp. 3-31.

[12] RoboCup Singapore, "CoSpace Robotics," 2015. [Online]. Available: http://cospacerobot.org/. [Accessed 27 June 2018].

[13] FIRST, "FIRST Inspires," 2018. [Online]. Available: https://www.firstinspires.org/. [Accessed 27 June 2018].

[14] New Energy Industrial Technology Development Organisation (NEDO), "World Robot Summit," 2018. [Online]. Available: http://worldrobotsummit.org/en/about/. [Accessed 27 June 2018].

[15] R. Sheh, A. Jacoff, A.-M. Virts, T. Kimura, J. Pellenz, S. Schwertfeger and J. Suthakorn, "Advancing the state of urban search and rescue robotics through the robocuprescue robot league competition," in *Field and Service Robotics*, 2012.

[16] RoboCup Federation, "RoboCup Junior," 2018. [Online]. Available: http://junior.robocup.org/. [Accessed 27 June 2018].

[17] Arduino, "Arduino," 2018. [Online]. Available: https://www.arduino.cc/. [Accessed 27 June 2018].

[18] Raspberry Pi Foundation, "Raspberry Pi," 2018. [Online]. Available: https://www.raspberrypi.org/. [Accessed 27 June 2018].

[19] R. Sheh, "The Open Academic Robot Kit," 2018. [Online]. Available: http://www.oarkit.org/. [Accessed 27 June 2018].

[20] R. Sheh, A. Eguchi, H. Komsuoglu and A. Jacoff, "The Open Academic Robot Kit," in *Robotics in STEM Education*, Springer, Cham, 2017, pp. 85-100.

[21] Robotis, "TurtleBot 3," 2018. [Online]. Available: http://www.robotis.us/turtlebot-3/. [Accessed 27 June 2018].

[22] INRIA, "Poppy Project," 2016. [Online]. Available: https://www.poppy-project.org/en/. [Accessed 27 June 2018].

[23] Vorpal Robotics LLC, "Vorpal Robots," 2018. [Online]. Available: https://www.vorpalrobotics.com/. [Accessed 27 June 2018].

[24] R. Ma, L. Odhner and A. Dollar, "A Modular, Open-Source 3D Printed Underactuated Hand," in *IEEE International Conference on Robotics and Automation (ICRA),* Karlsruhe, Germany, 2013.

[25] *ASTM Standard E2804-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Obstacles: Stairs/Landings,* West Conshohocken, PA: ASTM International, www.astm.org, 2011.

[26] *ASTM Standard E2826-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Continuous Pitch/Roll Ramps,* West Conshohocken, PA: ASTM International, www.astm.org, 2011.

[27] *ASTM Standard E2827-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Terrains: Crossing Pitch/Roll Ramps,* West Conshohocken, PA: ASTM International, www.astm.org, 2011.

[28] "Subcommittee E54.09 on Response Robots," ASTM International, 2019. [Online]. Available: https://www.astm.org/COMMIT/SUBCOMMIT/E5409.htm. [Accessed April 2019].

[29] A. Jacoff, E. Messina, H.-M. Huang, A. Virts, A. Downs, R. Norcross and R. Sheh, "Guide for Evaluating, Purchasing, and Training with Response Robots Using DHS-NIST-ASTM International Standard Test Methods," 2013. [Online]. Available:

https://www.nist.gov/sites/default/files/documents/el/isd/ks/DHS_NIST_ASTM_Robot_Test_Methods-2.pdf. [Accessed 27 June 2018].

[30] *ASTM Standard E2802-11 Standard Test Method for Evaluating Emergency Response Robot Capabilities: Mobility: Confined Area Obstacles: Hurdles,* West Conshohocken, PA: ASTM International, www.astm.org, 2011.

[31] *ASTM Standard E2992/E2992M-17 Standard Test Method for Evaluating Response Robot Mobility: Traverse Sand Terrain,* West Conshohocken, PA: ASTM International, www.astm.org, 2-17.

[32] *ASTM Standard E2991/E2991M-17 Standard Test Method for Evaluating Response Robot Mobility: Traverse Gravel Terrain,* West Conshohocken, PA: ASTM International, www.astm.org, 2017.

[33] *ASTM Standard E2566-17a Standard Test Method for Evaluating Response Robot Sensing: Visual Acuity,* West Conshohoken, PA: ASTM International, www.astm.org, 2017.

Automated fitting of thermogravimetric analysis data

Morgan C. Bruns[1] | **Isaac T. Leventon**[2]

[1]Department of Mechanical Engineering, Virginia Military Institute, Lexington, Virginia

[2]Fire Research Division, National Institute of Standards and Technology, Gaithersburg, Maryland

Correspondence
Morgan C. Bruns, Department of Mechanical Engineering, Virginia Military Institute,
710 Nichols Hall, Lexington, VA 24450.
Email: morgan.chase.bruns@gmail.com

Summary

A novel methodology has been developed for extracting pyrolysis kinetic parameters from thermogravimetric analysis (TGA) data. The development of this methodology is motivated by a need to automate the determination of material properties for use in fire models. The algorithm with which the methodology is implemented is described. Aside from being fully automated, the resultant script has the advantage of being efficient—a full set of kinetic parameters is provided in less than 1 second. The script is verified against manufactured TGA data for one and two reaction mechanisms and the effects of reaction peak width and the distance between reaction peaks is examined. Validation is accomplished by applying the script to TGA data for Nylon 6,6, a flexible polyurethane foam, and polyvinyl chloride. The resultant kinetic parameters are tabulated, and plots of the actual and predicted TGA data show that the algorithm is quite effective for one, two, and three reaction mechanisms.

KEYWORDS

fire modeling, kinetics, nylon 6,6, polyurethane foam, polyvinyl chloride, pyrolysis, TGA

1 | INTRODUCTION

Computational fire models have proven to be effective at predicting the spread of heat and smoke in a wide range of building fire scenarios. However, such models still generally require user input describing the actual fire that is generating the heat and smoke. Computational predictions of flame spread and fire growth require somewhat detailed models of condensed phase physics, and a number of condensed phase pyrolysis models have been developed.[1-3] Such models have proven effective at modeling the burning rate of small slabs, but their application to flame spread calculations is more limited. Part of the problem is that these pyrolysis models require the specification of a large number of material properties. Furthermore, there are many different flammable materials that need to be considered in fire scenarios. Substantial progress in applying computational fire models to predict flame spread could be achieved by the development of both a streamlined methodology for characterizing the thermophysical properties of flammable materials as well as the creation of a publicly available database of such material properties.

One successful approach for characterizing materials is based on performing a number of milligram-scale and bench scale tests such as thermogravimetric analysis (TGA), differential scanning calorimetry (DSC), and the controlled atmosphere pyrolysis apparatus.[4] An example of a public database of material properties is found in the Validation Guide of the Fire Dynamics Simulator (FDS).[5] Currently, an international effort is underway to further develop experimental and modeling tools such as these in order to “advance predictive fire modelling”.[6] The present manuscript presents work based on a coordinated effort to develop a comprehensive database of material properties for use in fire models. This database will include both raw data from a suite of milligram-scale tests and a list of properties for each material that could be directly used as inputs for a fire model such as FDS. In order to generate these tables of material properties, automated computational scripts are required to robustly analyze raw data from small-scale tests. In this paper, we present such a script for calibrating a generalized pyrolysis kinetic model to TGA data. Previous work has looked at optimization algorithms[7] and Markov Chain Monte Carlo methods[8] for fitting TGA data, but these approaches are not fully automated for general multistep reaction schemes and can be relatively computationally expensive.

Leventon, Isaac; Bruns, Morgan. ”Automated Fitting of Thermogravimetric Analysis Data.” Paper presented at Interflam 2019, London, UK.
July 01, 2019 - July 03, 2019.

2 | THEORY

In TGA, the mass of a small sample of material is measured while heated according to a prescribed temperature program. Typically, the sample will be heated at a constant temperature ramp rate. If the sample mass and heating rate are sufficiently small, then the temperature and composition throughout the sample are approximately uniform. As the material is heated, chemical bonds are broken producing smaller molecules. Eventually, the products of pyrolysis become small enough to vaporize, and mass is lost from the system. Models of this process typically take the form of a system of reactions with unimolecular Arrhenius kinetics.

2.1 | Independent Unimolecular Reactions

In the following, focus is limited to systems of independent reactions. This generalized model could account for (a) a system of parallel reactions or (b) a system of series reactions in which the subsequent reactions occur at different temperatures. To begin, consider a system of N_r unimolecular reactions of the form

$$R_i \xrightarrow{k_i} \nu_i C + (1-\nu_i)G, i = 1, \ldots, N_r, \tag{1}$$

where R_i is the label for the reactant material, C is the label for the condensed phase products of the reaction, k_i is the reaction rate constant, ν_i is the residual mass fraction of the reaction, and G is the label for the gas species which escapes the sample. For the purposes of the following analysis, it is not necessary to distinguish between the condensed phase and gas phase products of the reactions. Additionally, consideration will be limited to constant heating rate TGA experiments in which the sample temperature increases at a constant rate, β. An Arrhenius model is assumed for the temperature dependence of the rate constant such that, for each reaction

$$k_i = \left(\frac{A_i}{\beta}\right) \exp\left(-\frac{E_i}{RT}\right), \tag{2}$$

where A_i is the pre-exponential, E_i is the activation energy, and R is the gas constant. Note that the heating rate has been directly absorbed into the rate constant in order to simplify the notation in the following analysis. Consequently, k_i is not strictly speaking the Arrhenius rate constant, but the kinetic pair (A_i, E_i) are the true Arrhenius pre-exponential and activation energy, respectively.

For the kinetic model described by Equations (1), the temperature rate of change of the mass of component i, m_i, is governed by

$$m_i' \equiv \frac{\mathrm{d}m_i}{\mathrm{d}T} = -m_i k_i, \tag{3}$$

along with the initial condition $m_i(T_0) = m_{0,\,i}$ where T_0 is the initial temperature of the TGA experiment (or any temperature prior to the onset of the reaction). The total sample mass of the sample is simply the sum of the individual component masses. The total mass loss rate must account for the generation of residual solid products of the reactions so that

$$m' \equiv \frac{\mathrm{d}m}{\mathrm{d}T} = \sum\nolimits_i (1-\nu_i) m_i'. \tag{4}$$

2.2 | Nondimensional form and approximate solution

It is convenient to nondimensionalize Equation (3), and the resultant nondimensionalization leads to an approximate solution which is valid for most materials. The peak mass loss rate may be found by differentiating Equation (3) with respect to temperature and setting the result equal to zero. Analysis of this peak condition yields the following results. Using $T_{\mathrm{p},\,i}$, $m_{\mathrm{p},\,i}$, and $m'_{\mathrm{p},i}$ to denote the temperature, mass, and temperature derivative of the mass corresponding to that peak, a characteristic width of the peak may be defined as

$$\Delta T_i \equiv \frac{-m_{\mathrm{p},i}}{m'_{\mathrm{p},i}} = \frac{RT^2_{\mathrm{p},i}}{E_i}, \tag{5}$$

where the second equality in Equation (5) is determined from analysis of the peak equation. The kinetic equations can be recast in terms of the peak temperature and the peak width parameters which are related to the Arrhenius parameters through

$$E_i = \frac{RT^2_{\mathrm{p},i}}{\Delta T_i}. \tag{6}$$

$$A_i = \frac{\beta}{\Delta T_i} \exp\left(\frac{T_{\mathrm{p},i}}{\Delta T_i}\right). \tag{7}$$

A simplified form of the kinetic equations may be obtained using the following nondimensionalization

$$\mu_i \equiv \frac{m_i}{m_{0,i}}. \tag{8}$$

$$\theta_i \equiv \frac{T - T_{\mathrm{p},i}}{\Delta T_i}. \tag{9}$$

The nondimensional kinetic equation for component i becomes

$$\mu_i' \equiv \frac{\mathrm{d}\mu_i}{\mathrm{d}\theta_i} = -\mu_i \exp\left(\frac{\theta_i}{\xi_i \theta_i + 1}\right), \tag{10}$$

with the boundary condition $\mu_i(\theta_i \to -\infty) = 1$ where $\xi_i \equiv \Delta T_i / T_{\mathrm{p},\,i}$ may be thought of as a shape parameter for the mass loss peak.

For the limiting case in which $\xi_i \rightarrow 0$, Equation (10) has the exact solution:

$$\mu_i = \exp[-\exp(\theta_i)], \tag{11}$$

or, in dimensional form,

$$m_i = m_{0,i}\exp\left[-\exp\left(\frac{T-T_{\mathrm{p},i}}{\Delta T_i}\right)\right]. \tag{12}$$

Note that Equation (12) predicts that that mass at the peak rate of temperature change is $m_{0,\ i}e^{-1}$ which is the same approximation derived by other authors.[9] Furthermore, the assumption of letting $\xi_i \rightarrow 0$ is equivalent to $RT_{\mathrm{p},\ i} \ll E_i$ which is the same assumption used to arrive at this approximation.

2.3 | Reaction Peak Analysis

Equation (12) may be used in conjunction with TGA data to get estimates of the kinetic parameters. The approach developed in the following makes use of the fact that if the peak temperatures, $T_{\mathrm{p},\ i}$, are known from the data, it is possible to use derivatives of Equation (12) to obtain estimates of ΔT_i and the total mass lost from the sample due to the reaction, Δm_i, from the data. Note that the total mass change associated with a reaction is

$$\Delta m_i \equiv m_{0,i}(1-\nu_i). \tag{13}$$

Taking the first three derivatives with respect to temperature of Equation (12) results in:

$$d_{1,i} \equiv (1-\nu_i)m_i' = \frac{\Delta m_i}{\Delta T_i}g_i\exp(g_i), \tag{14}$$

$$d_{2,i} \equiv (1-\nu_i)m_i'' = \frac{\Delta m_i}{\Delta T_i^2}g_i\exp(g_i)(g_i+1), \tag{15}$$

$$d_{3,i} \equiv (1-\nu_i)m_i''' = \frac{\Delta m_i}{\Delta T_i^3}g_i\exp(g_i)\left(g_i^2+3g_i+1\right), \tag{16}$$

where

$$g_i \equiv -\exp\left(\frac{T-T_{\mathrm{p},i}}{\Delta T_i}\right). \tag{17}$$

Note that the newly defined d-variables in Equations (14) to (16) represent the contribution of each component to the first three temperature derivatives of the total mass. Thus,

$$m' = \sum\nolimits_i d_{1,i}. \tag{18}$$

$$m'' = \sum\nolimits_i d_{2,i}. \tag{19}$$

$$m''' = \sum\nolimits_i d_{3,i}. \tag{20}$$

At the peak temperature, $g_i = -1$, and so Equations (14) to (16) yield

$$d_{1,i}\left(T_{\mathrm{p},i}\right) = -\frac{\Delta m_i}{e\Delta T_i}. \tag{21}$$
$$d_{2,i}\left(T_{\mathrm{p},i}\right) = 0. \tag{22}$$

$$d_{3,i}\left(T_{\mathrm{p},i}\right) = \frac{\Delta m_i}{e\Delta T_i^3}. \tag{23}$$

Equation (22) shows that the peak condition is in fact being satisfied. The solution of Equations (21) and (23) in gives

$$\Delta T_i = \sqrt{-\frac{d_{1,i}\left(T_{\mathrm{p},i}\right)}{d_{3,i}\left(T_{\mathrm{p},i}\right)}}, \tag{24}$$

$$\Delta m_i = -ed_{1,i}\left(T_{\mathrm{p},i}\right)\Delta T_i, \tag{25}$$

where using Equations (18) and (20),

$$d_{1,i}\left(T_{\mathrm{p},i}\right) = m'\left(T_{\mathrm{p},i}\right) - \sum\nolimits_{j\neq i} d_{1,j}\left(T_{\mathrm{p},i}\right). \tag{26}$$

$$d_{3,i}\left(T_{\mathrm{p},i}\right) = m'''\left(T_{\mathrm{p},i}\right) - \sum\nolimits_{j\neq i} d_{3,j}\left(T_{\mathrm{p},i}\right). \tag{27}$$

In the following section, an algorithm is presented for using Equations (24) to (27) in conjunction with TGA data to obtain a set of kinetic parameters.

2.4 | Description of algorithm

In order to use Equations (24) to (27), it is necessary to have the temperature derivatives of the TGA data. That is, given the set of TGA data points for (T, m), what are the corresponding values of m', m'', and m''' at each temperature? It was found that the Savitzky-Golay filter[10] proved to be effective in estimating the derivatives of noisy TGA data. Application of the Savitzky-Golay filter requires two parameters: (a) the order of the polynomial fit and (b) the number of data points used in the fit. A quadratic fit was used for the first temperature derivative, a cubic fit for the second derivative, and a quartic fit was used for the third derivative data. The number of data points used in the fit was chosen to be enough to cover a temperature window of 10 K for the first temperature derivative, 20 K for the second derivative, and 40 K for the third derivative. These parameters were determined by trial and error, and future work will examine optimizing the choice of these parameters.

Once the first 3 derivatives of the TGA mass data were determined, it is necessary to determine the number of reactions. Clear reaction peaks correspond to points at which $m'' = 0$ and $m''' > 0$. All such peaks are easily obtained from the filtered temperature derivatives. Only peaks in which the mass loss rate was greater than 15% of the maximum peak mass loss rate were identified as distinct reactions. If two reactions occur over similar temperature ranges, it is often difficult to observe a distinct peak in the mass loss rate. However, in some cases the presence of a shoulder in the peak is evidence of a partially obscured reaction. Partially obscured reactions were found by locating points where m'' is small and $m''' = 0$.

The preceding paragraph presents two criteria for identifying reaction peaks. Each of these points corresponds to a temperature at which the peak occurs. So at this point in the algorithm, all that is stored is a set of N_r temperatures, $T_{p,\,i}$ for $i = 1, \ldots, N_r$. The next step is to determine the kinetic parameters for each of these reactions.

For the single reaction case, the total volatile mass lost from the reaction is equal to the total mass lost by the TGA sample, or

$$\Delta m_1 = \Delta m_{\mathrm{tot}} \equiv m_0 - m_{\mathrm{f}}, \tag{28}$$

where m_0 and m_{f} are the initial and final masses of the sample. With only one reaction, Equation (26) gives

$$d_{1,1}(T_{\mathrm{p},1}) = m'(T_{\mathrm{p},1}), \tag{29}$$

and so, upon substitution into and rearrangement of Equation (25),

$$\Delta T_1 = -\frac{\Delta m_{\mathrm{tot}}}{e m'(T_{\mathrm{p},1})}. \tag{30}$$

If more than one peak is found, the process is complicated since the mass loss rate signals can overlap. An iterative approach has been developed in which an initial guess of all Δm_i and ΔT_i are used to compute Equations (26) and (27), and then the resultant values are used in Equations (24) and (25) to find updated values for these kinetic parameters. The process is repeated until convergence is achieved.

An important correction must be employed to insure that $\sum_i \Delta m_i = \Delta m_{\mathrm{tot}}$. Using Equations (24) and (25) for each reaction separately does not guaranty that this condition will be satisfied. In order, to identify a mass conserving mechanism an intermediate solution ΔT_i^* is found from Equation (24). The mass conserving peak widths would be those which satisfy:

$$\Delta m_{\mathrm{tot}} = \sum_i -e d_{1,i}(T_{\mathrm{p},i}) \Delta T_i. \tag{31}$$

The projection of the intermediate solution onto the space of solutions satisfying Equation (31) is found by:

$$\mathbf{\Delta T} = \mathbf{\Delta T}^* - \left(\frac{\mathbf{\Delta T}^* \cdot \mathbf{a} - \Delta m_{\mathrm{tot}}}{\mathbf{a} \cdot \mathbf{a}} \right) \mathbf{a}, \tag{32}$$

where $\mathbf{\Delta T}^*$ and $\mathbf{\Delta T}$ are simply the vectors formed from the temperature widths for the intermediate and mass conserving cases, and $\mathbf{a}$ is the vector whose elements are $a_i \equiv -e d_{1,\,i}(T_{\mathrm{p},\,i})$. The mass widths for an iteration are then found from the mass conserving temperature widths used in Equation (25).

The entire algorithm for finding the kinetic parameters of TGA data indicating several reactions is summarized as follows:

1 Estimate reaction mass changes using $\Delta m_i = \Delta m_{\mathrm{tot}}/N_r$ for each i.
2 Compute corresponding estimates of peak widths using $\Delta T_i = -\Delta m_i / e m^{'}(T_{\mathrm{p},\,i})$ for each i.
3 Compute $d_{1,\,i}$ and $d_{3,\,i}$ for each i using the previously determined values for Δm_i and ΔT_i.
4 Find initial estimates ΔT_i^* for each i using Equation (24).
5 Calculate mass conserving temperature widths using Equation (32).
6 Calculate mass conserving reaction mass changes using Equation (25).
7 Repeat steps 3 through 6 until Δm_i and ΔT_i converge.

It was found that even for the most complex scenarios considered that convergence was achieved after approximately 10 to 20 iterations making the algorithm extremely efficient. In the next two sections, the algorithm described in this section will be verified against manufactured data and validated against TGA data for several materials. In all cases, the algorithm requires less than 1 second of CPU time on a typical laptop computer to provide the complete set of kinetic parameters.

3 | VERIFICATION

It is possible to verify the algorithm described in the preceding section by testing it against numerically generated TGA data based on an assumed kinetic model. In this process, TGA data is manufactured by simulating a solution of Equation (4) using assumed parameter values. The algorithm was then applied to this manufactured data. The resultant calibrated kinetic parameters can be compared to the specified value, and predictions using the calibrated kinetic parameters can be compared to the manufactured data. In all cases, a heating rate of 10 K/minute was used along with an initial sample mass of $m_0 = 1$.

3.1 | One reaction

Three single-reaction cases have been considered. Each of these three cases specify a peak reaction rate temperature at 650 K, and the sample was assumed to fully volatilize so that $m_{\mathrm{f}} = 0$ and $\Delta m = 1$. In order to examine the effectiveness of the algorithm to handle a variety of data, three different characteristic peak widths were considered: $\Delta T = 10$ K, $\Delta T = 20$ K, and $\Delta T = 40$ K. The specified and calibrated parameters for these three cases are shown in Tables 1 to 3. Plots of the TGA mass and mass loss rate for the three cases are shown in Figure 1. It is clear that the quality of the calibrated reaction decreases with increasing reaction width. This observation is likely a consequence of the fact that the fitting algorithm is based upon an approximate solution of the kinetic equations that assumes a small value of $\xi \equiv \Delta T / T_{\mathrm{p}}$. As the value of ΔT increases, so does the value of ξ, and thus the validity of the approximation decreases.

TABLE 1 Kinetic parameters for the single-reaction verification case with ΔT = 10 K

Kinetic parameter	Specified value	Calibrated value
T_p (K)	650	649.4
ΔT (K)	10	9.99
ξ	0.01538	0.01539
ln[A (s^{-1})]	60.91	60.90
E (kJ/kmol)	351.3×10^3	350×10^3

TABLE 2 Kinetic parameters for single-reaction verification case with ΔT = 20 K

Kinetic parameter	Specified value	Calibrated value
T_p (K)	650	649.4
ΔT (K)	20	19.07
ξ	0.03077	0.02935
ln[A (s^{-1})]	27.71	29.34
E (kJ/kmol)	175.6×10^3	184.1×10^3

TABLE 3 Kinetic parameters for single-reaction verification case with ΔT = 40 K

Kinetic parameter	Specified value	Calibrated value
T_p (K)	650	649.4
ΔT (K)	40	36.
ξ	0.06154	0.05563
ln[A (s^{-1})]	10.77	12.59
E (kJ/kmol)	87.8×10^3	97.1×10^3

3.2 | Two reactions

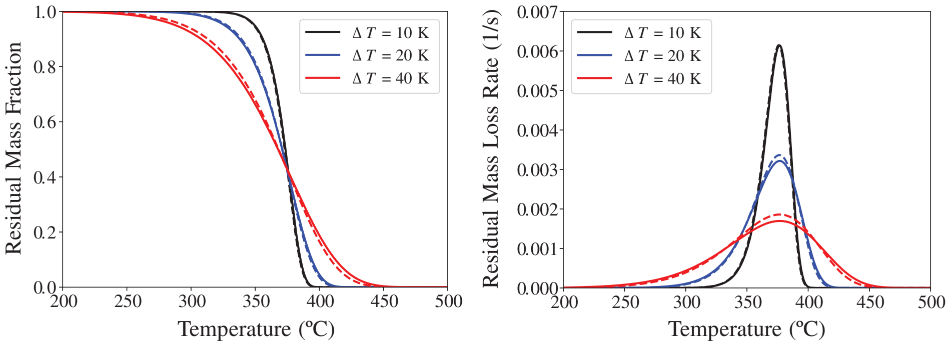
Three 2-reaction cases were also considered. For these cases, the effect of overlapping peaks was studied. For these scenarios, both reactions are assigned a characteristic peak width of 15 K. The normalized mass change associated with the first reaction is Δm_1 = 0.5, and that of the second reaction is Δm_2 = 0.3. In all cases, the midpoint between the two reaction peaks is 650 K. The difference between the cases is the distance between the peak temperatures. Three peak temperature differences were considered: 160, 80, and 40 K. The specified and calibrated parameters for these three cases are shown in Tables 4 to 6. Plots of the results of these three verification cases are shown in Figure 2. It is apparent that the algorithm captures the TGA signal well even as the peaks move close together. Although the plateau in the mass signal is overestimated, the mass loss rates are well captured—this is a consequence of the algorithm favoring mass loss rate matching over mass matching.

TABLE 4 Kinetic parameters for two-reaction verification case with $T_{p,2} - T_{p,1}$ = 160 K

Kinetic parameter	Specified value	Calibrated value
$T_{p,1}$ (K)	570	569.7
$T_{p,2}$ (K)	730	729.6
ΔT_1 (K)	15	14.12
ΔT_2 (K)	15	15.25
ξ_1	0.02632	0.02478
ξ_2	0.02055	0.02090
Δm_1	0.5	0.4874
Δm_2	0.3	0.3126
ln[A_1 (s^{-1})]	33.50	35.91
ln[A_2 (s^{-1})]	44.17	43.34
E_1 (kJ/kmol)	180.1×10^3	191.1×10^3
E_2 (kJ/kmol)	295.4×10^3	290.2×10^3

4 | VALIDATION

The manufactured TGA data considered above represent an idealization of the complex processes actually occurring during the pyrolysis



(A) (B)

FIGURE 1 Specified and calibrated TGA (A) normalized mass and (B) normalized mass loss rates for single-reaction verification cases. The solid lines represent the manufactured data and the dashed lines represent model predictions

TABLE 5 Kinetic parameters for two-reaction verification case with $T_{p,2} - T_{p,1}$ = 80 K

Kinetic parameter	Specified value	Calibrated value
$T_{p,1}$ (K)	610	609.7
$T_{p,2}$ (K)	690	689.7
ΔT_1 (K)	15	14.20
ΔT_2 (K)	15	15.26
ξ_1	0.02459	0.02329
ξ_2	0.02174	0.02212
Δm_1	0.5	0.4865
Δm_2	0.3	0.3135
ln[A_1 (s^{-1})]	36.17	38.49
ln[A_2 (s^{-1})]	41.50	40.69
E_1 (kJ/kmol)	206.2 × 10^3	217.6 × 10^3
E_2 (kJ/kmol)	263.9 × 10^3	259.3 × 10^3

TABLE 6 Kinetic parameters for two-reaction verification case with $T_{p,2} - T_{p,1}$ = 40 K

Kinetic parameter	Specified value	Calibrated value
$T_{p,1}$ (K)	630	631.2
$T_{p,2}$ (K)	670	669.6
ΔT_1 (K)	15	14.10
ΔT_2 (K)	15	16.59
ξ_1	0.02381	0.02234
ξ_2	0.02239	0.02477
Δm_1	0.5	0.4590
Δm_2	0.3	0.3410
ln[A_1 (s^{-1})]	37.50	40.32
ln[A_2 (s^{-1})]	40.17	35.77
E_1 (kJ/kmol)	220.0 × 10^3	234.9 × 10^3
E_2 (kJ/kmol)	248.8 × 10^3	224.7 × 10^3

TABLE 7 Calibrated kinetic parameters for Nylon 6,6

Kinetic parameter	Reaction 1
T_p (K)	716.3
ΔT (K)	22.11
Δm	0.9754
ξ	0.03087
ln[A (s^{-1})]	27.50
E (kJ/kmol)	192.9 × 10^3

of real materials. It is therefore important to assess the validity of the fitting algorithm against real TGA data.

In this work, TGA experiments were conducted on three polymers: Polyamide 6,6 (Nylon 6,6), a flexible polyurethane (PU) foam, and polyvinyl chloride (PVC). These materials are widely used in construction, residential, and transportation applications, and, collectively, they present a diverse range of decomposition behaviors (eg, single or multiple reaction peaks, discrete or overlapping reactions, and decomposition across a wide temperature range).

TGA experiments were conducted on these three materials in a Netzsch STA 449 F1 Jupiter. This apparatus continuously measures mass (using a microbalance with a 0.025 μg precision) and temperature (using an S-type thermocouple positioned directly beneath the sample crucible) of samples as they are heated through a well-defined temperature program in an anaerobic environment. A temperature calibration was conducted as per the manufacturer's recommendations[11] (using a set of 6 pure metals, with melting points between 156.6 and 961.8°C) to provide a relation between measured and actual sample temperature. The calibration was performed using the same crucible type, heating rate, and gaseous environment as was used during thermal analysis experiments on the polymeric samples. All TGA experiments were conducted within 3 months of this calibration.

The temperature program used for TGA experiments included an initial isotherm at an elevated temperature (below 100°C), during

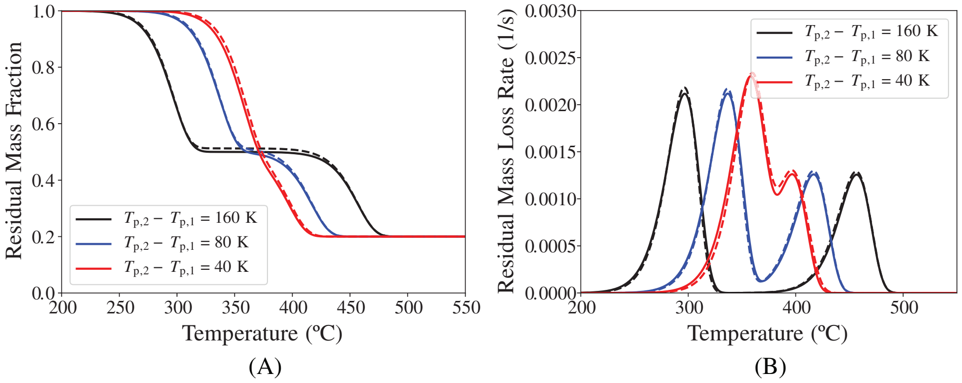


FIGURE 2 Specified and calibrated TGA (A) normalized mass and (B) normalized mass loss rate for two-reaction verification cases. The solid lines represent the manufactured data and the dashed lines represent model predictions

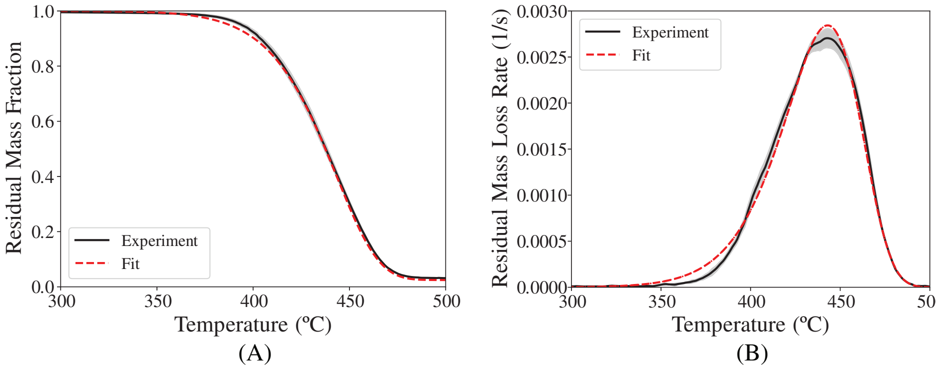


FIGURE 3 Experimental and calibrated TGA (A) normalized mass and (B) normalized mass loss rate for Nylon 6,6. The gray area represents ± two SDs about the average mass

TABLE 8 Calibrated kinetic parameters for polyurethane foam

Kinetic parameter	Reaction 1	Reaction 2
T_p (K)	562.7	648.5
ΔT (K)	14.50	13.69
Δm	0.2511	0.7280
ξ	0.02577	0.02112
$\ln[A\ (s^{-1})]$	34.34	42.95
E (kJ/kmol)	181.5×10^3	255.3×10^3

TABLE 9 Calibrated kinetic parameters for polyvinyl chloride

Kinetic parameter	Reaction 1	Reaction 2	Reaction 3
T_p (K)	568.5	731.7	588.1
ΔT (K)	12.15	22.39	9.62
Δm	0.4200	0.2238	0.1999
ξ	0.02138	0.03060	0.01636
$\ln[A\ (s^{-1})]$	42.49	27.78	57.06
E (kJ/kmol)	221.1×10^3	198.8×10^3	298.8×10^3

which time the chamber was continuously purged with nitrogen. This ensured that the system was completely free of oxygen and that any residual moisture in samples was removed prior to dynamic heating and thermal decomposition. Following this conditioning period, samples were heated at a constant rate of 10 K/min up to 700 or 900°C. Throughout this program, the test chamber was continuously purged with ultra-high purity (UHP) nitrogen at 50 mL/min. All tests were conducted in open alumina crucibles.

At the start of each day of testing, a baseline test was performed in which an empty alumina crucible was subjected to the same heating program as was used during the thermal analysis experiments. This baseline history (mass vs temperature) was subtracted from the corresponding data obtained during experiments on the polymeric samples. All TGA measurement data presented in this work has been baseline-corrected in this manner. All samples were stored in a desiccator (in the presence of Drierite) for a minimum of 48 hours prior to testing. Immediately before testing, samples were removed from the desiccator, placed into alumina test crucibles, and weighed using a Mettler M3 analytical balance.

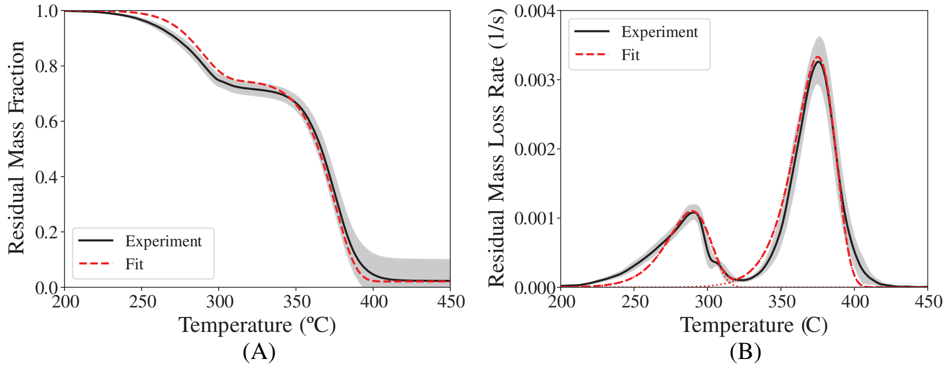


FIGURE 4 Experimental and calibrated TGA (A) normalized mass and (B) normalized mass loss rate for polyurethane foam. The gray area represents ± two SDs about the average mass

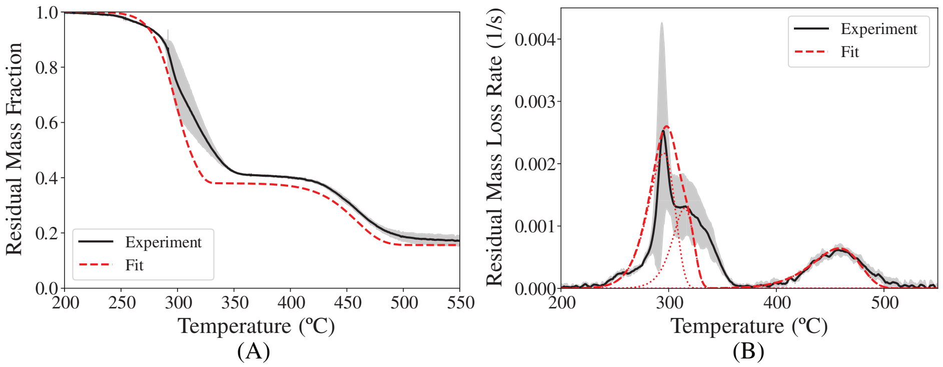


FIGURE 5 Experimental and calibrated TGA (A) normalized mass and (B) normalized mass loss rate for polyvinyl chloride. The gray area represents ± two SDs about the average mass

4.1 | Polyamide 6,6 (Nylon 6,6)

Polyamide 6,6 (Nylon 6,6) samples were obtained from Goodfellow (Reference number AM323100) in the form of injection molded slabs, which were then cryogenically ground to obtain a powder for testing. For each test, between 4.5 and 5.5 mg of this powder was weighed and then pressed flat into the base of an alumina test crucible. The temperature program included a 20-minute-long isotherm at 27°C followed by heating at a constant rate of 10 K/min to 900°C. Three replicate tests were performed, and the average mass data was used by the algorithm to determine pyrolysis kinetics.

The calibrated kinetic parameters are listed in Table 7. Plots of the TGA data and the fit resulting from the calibrated parameters are provided in Figure 3 along with a gray area representing +/− two standard deviations of the mass data about the average mass. Although the initial onset temperature and the peak mass loss rate are slightly overpredicted, the predicted TGA agrees quite well with the experimental data.

4.2 | Flexible PU Foam

PU foam samples were produced by Innocor Foam Technologies to meet the specifications of a "Standard Polyurethane Foam Substrate" described in ASTM-D3574-08.[12] This foam was purchased in the form of 8 cm thick slabs. Samples used for testing were cut from the center of these slabs into small, cylindrical pieces, between 3.0 and 5.0 mg in mass. These foam pieces were then compressed for a minimum of 72 hours before being pressed flat in the center of an alumina test crucible using a steel reshaping tool, just prior to testing. The temperature program included a 20-minute-long isotherm at 75°C followed by heating at a constant rate of 10 K/min to 700°C. Five replicate TGA tests were performed, and the average results were used to estimate kinetic parameters. The calibrated kinetic parameters are provided in Table 8. A plot of the experimental data along with the fit parameters is given in Figure 4. Again, the results predicted using the calibrated kinetic parameters agree quite well with the TGA data. In fact the predicted peak mass loss rates for both reactions are approximately equal to the experimental values.

4.3 | Polyvinyl Chloride

PVC samples were obtained from Interstate Plastics (manufactured by Vycom plastics: Type 1 PVC) in the form of 6-mm thick slabs. For each test, small, flat pieces between 4.5 and 5.5 mg in mass, were carefully cut from these slabs and placed in the center of an alumina test crucible. The temperature program included a 30-minute-long isotherm at 40°C followed by heating at a constant rate of 10 K/minutes to 700°C. Six replicates were performed, and the average TGA signal was used to calibrate the kinetic parameters. The algorithm predicts a three-reaction mechanism, and the calibrated kinetic parameters are tabulated in Table 9. The data and corresponding fits are shown in Figure 5. It is apparent that the fit is not as good as was obtained for the other two materials. This can be attributed to the relative noisiness of the experimental data along with the fact that the first two reactions are relatively close to one another. However, the peak mass loss rates for all three reactions is very well captured by the calibrated kinetic model. The adequacy of this, or any fit, ultimately depends on the ability of the calibrated parameters to make accurate predictions of burning rate and flame spread in fire models. Such a study is beyond the scope of this paper.

Further validation of the algorithm has been performed for 11 different vegetative fuels.[13] These vegetative fuels typically exhibited two to three reaction peaks within close proximity and with considerable noise in the data. As can be confirmed in the reference, the automated fitting algorithm described in this paper performed well even in these more challenging scenarios.

5 | CONCLUSIONS

A novel methodology has been presented for finding the kinetic parameters of pyrolysis from TGA data. The advantages of this algorithm are that (a) it is fully automated requiring no interaction from the user, (b) it is computationally efficient in providing the kinetic parameters in less than 1 second, and (c) it has been verified and validated. The need for such a methodology is driven by a demand for material properties needed as inputs for fire models of flame spread.

Verification was performed by applying the algorithm to manufactured data for several one and two reaction pyrolysis mechanisms. As expected by the underlying theory, it was found that the algorithm performed better for narrower reactions. Additionally, slightly better fits were obtained for well-separated reaction peaks. For the two-reaction scenarios, it was found that the algorithm can sometimes fail to exactly predict mass plateaus between reactions. This is a consequence of the fact that the methodology focuses primarily on capturing the mass loss rate as opposed to the sample mass. A focus on the mass loss rate is justified since flammability is governed by the rate at which combustible volatiles are generated rather than the current amount of mass remaining in the sample. The results of the different verification cases demonstrate that even in situations in which the mass signal is not perfectly captured, the mass loss rate is quite accurate.

Three different materials were used to validate the algorithm: Nylon 66, PVC, and flexible PU foam. All TGA tests for these three materials were performed at a heating rate of 10 K/minutes in a nitrogen environment with the samples placed in an alumina crucible. These cases proved to be a rigorous test of the algorithm since the materials varied in the number of reactions predicted as well as the noise in the underlying data. Kinetic parameters for all three of these materials were determined using the automated algorithm, and the resultant model predictions agreed with the data nicely. A natural extension of this validation process is to test the derived kinetic parameters for TGA at different heating rates. It would be useful to explore the applicability of the calibrated kinetic parameters for making predictions of TGA at lower heating rates, and this remains an avenue for future work. However, a simple heat transfer analysis[14] indicates that the validity of assuming a spatially uniform sample temperature breaks down for heating rates much greater than 10 K/minutes for typical TGA sample sizes. Consequently, data obtained at larger heating rates would include transient effects from thermal (and possibly mass) transport rather than just pyrolysis kinetics. Such a conflation of physical processes will likely inhibit the effectiveness of the derived parameters to make predictions of full-scale scenarios when used in conjunction with separately derived transport properties.

Although the algorithm has been verified and validated to some extent, more validation is always desirable in order to find the limits of applicability of the algorithm. The results presented in this paper are quite promising, but it would be desirable to make better fits for broader peaked reactions as well as for overlapping reactions that occur at very similar temperatures. Furthermore, as was seen in the validation cases, different materials can produce different amounts of noise in the raw TGA data. It would therefore be helpful to make the algorithm more robust against varying degrees of noise. These improvements are primarily related to the very first step in the algorithm which uses the Savitzky-Golay filter to smooth the data and find sufficiently smoothed derivatives. There are several parameters involved in this process that need to be optimized by further study.

Finally, the true test of the effectiveness of this novel methodology lies in its effectiveness at providing parameters that accurately predict burning rate and flame spread in fire models such as FDS. Future work will therefore involve developing similar procedures to use other microscale data such as DSC and microscale combustion calorietery to estimate the other parameters needed in fire models. Once a more complete program of automation is achieved, it will be possible to validate the procedures by comparing fire model results with experimental flame spread data of real materials. Such a process will take significant effort, but the present paper provides a critical first step in this direction.

Disclaimer: The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

ORCID

Morgan C. Bruns https://orcid.org/0000-0001-6101-4383

REFERENCES

1. McGrattan KB, Hostikka S, McDermott R, Floyd J, Weinschenk C, Overholt K. *Fire Dynamics Simulator*, *Volume 1: Mathematical Model*. Vol 1. 6th ed. National Institute of Standards and Technology, NISTIR, Special Publication: Gaithersburg, MD; 2019:1018.
2. Lautenberger C, Fernanez-Pello C. Generalized pyrolysis model for combustible solids. *Fire Saf J*. 2009;44:819-839.
3. Stoliarov SI, Lyon RE. Thermo-kinetic model of burning. Technical Report DOT/FAA/AR-TN08/17, Federal Aviation Administration; 2008.
4. Stoliarov SI, Li J. Parameterization and validation of pyrolysis models for polymeric materials. *Fire Tech*. 2016;52:79-91.
5. McGrattan KB, Hostikka S, McDermott R, Floyd J, Weinschenk C, Overholt K. *Fire Dynamics Simulator: Technical Reference Guide*, *Volume 3: Validation*. Vol 1. 6th ed. Gaithersburg, MD: National Institute of Standards and Technology NISTIR, Special Publication; 2019:1018.
6. Brown A, Bruns M, Gollner M, et al. Proceedings of the first workshop organized by the IAFSS working group on measurement and computation of fire phenomena (MaCFP). *Fire Saf J*. 2018;101:1-17.
7. Bruns MC, Koo JH, Ezekoye OA. Population-based models of thermoplastic degradation: using optimization to determine model parameters. *Polymer Deg Stab*. 2009;94(6):1013-1022.
8. Bruns MC. Inferring and propagating kinetic parameter uncertainty for condensed phase burning models. *Fire Tech*. 2016;52(1):93-120.
9. McGrattan KB, Baum HR, Rehm RG, Hamins A, Forney GP. *Fire Dynamics Simulator: Technical Reference Guide*, *User's Guide*. 6th ed. National Institute of Standards and Technology, NISTIR, Special Publication: Gaithersburg, MD; 2019:1019.
10. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Analyt Chem*. 1964;36(8):1627-1639.
11. NETZSCH. *Software Manual (STA 449 F1 & F3) Temperature and Sensitivity Calibration*. NETZSCH Gerätebau GmbH: Selb, Germany; 2012.
12. ASTM. *D3574-08*. *Standard Test Methods for Flexible Cellular Materials—Slab, Bonded, and Molded Urethane Foams*. ASTM International: West Conshohocken, PA; 2008.
13. Leventon IT, Bruns MC. Thermal decomposition of vegetative fuels. Paper presented at: Proceedings of the 15th International Conference and Exhibition on Fire Science and Engineering (Interflam); 2019; London, UK.
14. Lyon RE, Safronava N, Senese J, Stoliarov SI. Thermokinetic model of sample response in nonisothermal analysis. *Thermochim Acta*. 2012; 545:82-89.

INFLUENCE OF DISPOSITIONAL AND SITUATIONAL FACTORS ON HUMAN PERCEPTIONS OF FIRE RISK

Justin W. Bonny [A] and Isaac T. Leventon *[B]

[A] Morgan State University, Department of Psychology;
1700 East Cold Spring Lane, Behavioral and Social Sciences Center, Room 431; Baltimore, MD, 21251; United States
[B] National Institute of Standards and Technology, Fire Research Division,
100 Bureau Drive; Building 224, Room A265; Gaithersburg, MD, 20899; United States

Keywords: Decision Making, Egress Modeling, Human Behavior in Fire, Psychological Traits, Risk Perception

ABSTRACT

At what point does an individual perceive a growing fire as dangerous and decide it is necessary to take action? Studies of human behavior during emergencies have observed that when presented with situational cues that a hazard may be present, humans can fail to act on this information in a timely manner. Models of human behavior in response to fire-related emergencies explicitly account for potential delays in identifying and acting on the presented risk of a fire. For example, the protective action decision model incorporates the tendency of individuals to notice an aberrant signal, but fail to recognize the cue as indicative of the need to evacuate, until it has become more salient. Past research has indicated that variations in the disposition of individuals, such as psychological traits, can also influence responses to emergencies. The present study examines the extent to which responses to images of growing room fires were influenced by situational and disposition factors. Participants judged whether words reflecting normalcy, risk, or protective action applied to images of room fires that varied in intensity. Psychophysical models of responses revealed, as the visual extent of the fires increased, deviation from normalcy words were first reliably judged to apply, then risk, and finally protective action, in line with models of human evacuation behavior. Individual differences in a specific dispositional trait, self-reported discomfort when exposed to sensory stimuli, was significantly related to performance: those who reported greater negative affect with sensory stimulation took longer to identify a growing fire as indicative of risk. The presence of moderate correlations between performance and other dispositional measures suggests future studies with greater statistical power may observe additional relations. Results of the current study lend support to the presence of a normalcy bias: even though participants noticed small fires and identified the scene as abnormal, they did not reliably view them as a risk until they grew larger. Furthermore, in line with evacuation models, once participants judged the fire cues to be a risk, fire size further increased before participants judged a protective action was applicable. The corroborating evidence in the present study suggests that psychophysical and observational data align with models of human evacuation behavior.

INTRODUCTION

A growing number of fire and life safety codes used around the world provide performance-based design options [1]. To ensure adequate provisions for occupant life safety, egress calculations are

Leventon, Isaac; Bonny, Justin. "Influence of Dispositional and Situational Factors on Human Perceptions of Fire Risk." Paper presented at Interflam 2019, London, UK. July 01, 2019 - July 03, 2019.

increasingly part of performance-based analyses [2]. The role of human behavior in egress calculations can vary from simple models that assume occupants will display identical responses with regard to pre-evacuation behavior (e.g., prescribed delay times in hydraulic based models [3]) to more comprehensive models that account for individual differences in occupant decision making processes (e.g., perception of hazards, and identification and assessment of risk [4–6]). Review papers of human behavior in fire stress the need to accurately assess pre-movement time (i.e., the time between first exposure to fire cues and movement towards a safe place) and have noted that it may be a more important element of required escape time than the time needed to move to a safe place [7,8]. Further, studies of real emergencies have shown that there is a connection between pre-evacuation time (delays in evacuation) and the number of fire deaths or injuries [8]. Unfortunately, multiple studies have highlighted that pre-evacuation behavior receives comparatively less attention than the actual process of evacuation behavior [9–12].

Pioneering efforts identified behaviors that are now recognized as key human responses to fire by surveying individuals after they had experienced a fire event and investigating the actions taken (as distinct from movement) when they encountered a fire in a building [13,14]. Later work would seek to better understand factors affecting human response (and related timing) in real fires including (but not limited to) occupancy [9,15], cues received [10], and occupant action prior to fire cue [11]. However, the fidelity of human behavior in response to fire cues incorporated within computational models that simulate egress is still limited [16] and further research is still needed to identify the underlying processes that produce these observable human behaviors in fire [12]. Including behavioral theories of human decision making regarding the timing of pre-evacuation actions and research linking influential factors (e.g., individual characteristics and environmental cues) to occupant decision making (e.g., identifying threats, perceiving risk, and selecting appropriate protective actions [17]) can improve these models. Greater understanding of how fire is identified as an imminent risk, a specific aspect of human decision making, could improve models and calculations of required safe egress time (RSET).

Models of occupant responses to fire emergencies suggest that specific factors can influence the perception of cues and assessment of the inherent risk of a situation. A specific model that has been applied to human behavior during emergencies, the Protective Action Decision Model (PADM), places an emphasis on the early processing of cues indicative of danger. According to this model, human responses to indicators of a present or imminent emergency depend on a series of processes where environmental cues can be perceived and considered when making decisions about how to respond [18]. When applied to fire-related emergencies, earlier processes are composed of perceiving environmental cues related to the emergency, with subsequent processes involving identifying whether these cues are indicative of risk, and later processes focusing on the decision making regarding taking protective action [17]. Studies of human behavior during emergency evacuations have observed that humans, at times, fail to perceive signs of a hazardous event as indicative of risk [19]. When applied to fire-related emergencies, the PADM explicitly incorporates biases to account for the potential delays in identifying and acting on the presented indicators of risk. Specifically, applications of the PADM to fire-related emergencies incorporate the tendency of individuals to fail to recognize aberrant signals (e.g., smoke) as abnormal; this is known as the normalcy bias [17]. This type of bias can shape initial perceptions as to whether changes in the environment are out of the ordinary and can lead to delays in responding to an emergency [17]. Previous research suggests that perception of cues as indicative of risk can be influenced by dispositional (e.g., psychological traits of the individual) and situational (e.g., environmental) factors [20]. This likely also extends to decision making regarding whether such cues warrant taking protective action [21]. When examining the impact of situational and dispositional factors on decision making, research that has applied the PADM to fire emergencies has predominantly used post hoc evaluations of human behavior during emergencies and interviews of individuals affected by emergencies as sources of data [19,20]. This leaves open questions regarding the extent to which perceptions and judgments when viewing fire cues are influenced by dispositional traits and situational factors.

In the present research, psychophysical models were used to assess at what point individuals viewing images of a developing fire reliably perceived a risk was present and required protective action. Participants viewed images selected from four full-scale compartment fires conducted in realistic, fully-furnished settings (two kitchen and two bedroom fires) and subsequently made judgments about the state of the room. Situational factors were varied by manipulating the stage in fire development that the image was taken from (pre-ignition to flashover) and by varying the type of judgment. Specifically, participants indicated whether a word prompt from select categories did or did not apply to the displayed image. Words were selected to generally reflect earlier versus later processes of the PADM, namely assessment of threat credibility, cues as indicators of a threat, and assessing need for protective action [22] – 'normal' or 'ordinary' (normalcy), 'danger' or 'emergency' (risk), 'evacuate' or 'flee' (protective action). The images ranged from rooms that did not contain flames or smoke, to rooms that had small fires that produced smoke, to rooms that had reached flashover. Psychometric functions were then fitted to collected responses to assess the point during fire growth at which participants reliably judged a word applied to the scene. To evaluate the extent to which variations in dispositional traits influenced judgments, participants also completed a series of questionnaires that assessed temperament (reactivity and self-regulation to changes in environment) and risk-taking tendencies. Individual differences in these factors have been observed to correlate with real-world behaviors such as risky driving [23] and pedestrian behavior [24], and recreational drug use [25]. Variations in temperament reflect differences in the way individuals react and self-regulate to changes in the environment, rooted in biological processes [26]. The focus of theories of temperament on the processes by which individuals respond to environmental factors suggests that individual differences in temperament may shape the ways in which individuals respond to fire cues. Furthermore, with risk taking involving actions that place an individual at an increased chance of experiencing harm [25], it was anticipated that risk-taking tendencies may also influence reactions to environmental indicators of an emergency. Correlational analyses examined whether individual differences in the psychometric curves were connected to variations in these psychological traits.

The anticipated impact of this research is two-fold. First, the novel application of psychophysical models to fire risk perception aimed to address concerns regarding past qualitative and rating-scale research when investigating the impact of situational factors [20]. It was predicted that judgments of growing room fires would align with earlier and later processes associated with human responses to emergencies: participants would initially detect environmental cues indicative of a deviation from normalcy, then interpret more intense fire cues as indicative of risk, and finally judge that the growing fire would require protective action. If successful, the novel approach can be systematically used to study the influence of multiple factors on fire risk perception. Second, the results can be used to inform models of risk perception in emergency scenarios. The development and evaluation of such comprehensive models could enable competent model users to predict key evacuee behaviors, and account for individual differences in dispositional traits, rather than directly prescribing such behaviors, a priori, for a given scenario.

METHOD

Participants

The responses from a total of 40 participants (5 males; age: mean, $M = 20.55$, standard deviation, $SD = 2.34$) were included in the present study. The majority of participants identified as being black ($N = 37$ including 4 Hispanic) with others identifying as white ($N = 2$ including 0 Hispanic) and of mixed race

(*N* = 1 including 0 Hispanic). Participants were recruited from undergraduate psychology courses at a mid-sized university in the Baltimore-Washington metropolitan area in the United States. Recruiting materials for the study emphasized that, as part of the study, participants would be viewing real images of room fires. For completing the study, participants received course credit. The research protocol was approved by an institutional review board and abided by the Declaration of Helsinki.

Hardware and Software

All study materials were presented on a laptop computer (39.6 cm diagonal screen) running the Windows 10 operating system. The judgment task was coded and presented via the Tobii Pro Lab (Tobii AB) software system. A custom Unity3D program was used to administer questionnaires.

Procedure

After providing written informed consent, participants were presented with the judgment task. Participants then completed three questionnaires and were provided with a debriefing form. Participants completed the study in about 35 to 45 minutes.

Judgment Task

Images presented during the judgment task were selected from four video clips of fires in realistically furnished rooms. Each video clip contained a room scene: specifically, two bedroom scenes and two kitchen scenes. A sequence of nine still images were extracted from each video clip such that the apparent size or intensity of the fire increased from no fire or smoke present, to low intensity (e.g., no flames, little smoke), and to high intensity (e.g., large flames, heavy smoke). Images were edited to occlude any labels that included timestamps or watermarks. The final set of images were 1024 x 768 pixels. Figure 1 plots each sequence of images (for each of the bedroom and kitchen scenes) viewed by participants in this study.

During the task, a set of words was presented with each still image (pseudorandom order). Words visually presented to participants during the judgment task were selected to reflect earlier versus later processes associated with human responses to emergencies. Specifically, two words reflecting normalcy (‘ordinary’, ‘normal’), risk (‘danger’, ‘emergency’), and protective action (‘escape’, ‘flee’) were identified during a literature search and pilot testing.

Images were presented full-sized on a black background followed by words presented on a black background in a white font. For the task, participants were presented with room fire images and were asked whether a presented word did or did not match the image. During each trial, participants were first presented with a fixation point (800 ms) that was replaced by an image of a room fire which remained on-screen for four seconds. Afterwards, a word was presented and remained on-screen until participants made a keypress response as to whether the word matched the previously presented image (‘Y’ key = does match, ‘N’ key = does not match) and the next trial began.

Blocks of trials that contained different words were presented in a pseudorandom order. At the beginning of each block, instructions were presented to the participants indicating the word that they were to use to judge each image (within-subject factor). Within each block, all room fire images were presented in a pseudorandom order (within-subject factor; 36 trials total per block; 216 total trials across blocks). Participants were randomly assigned to complete one of three different trial orders.

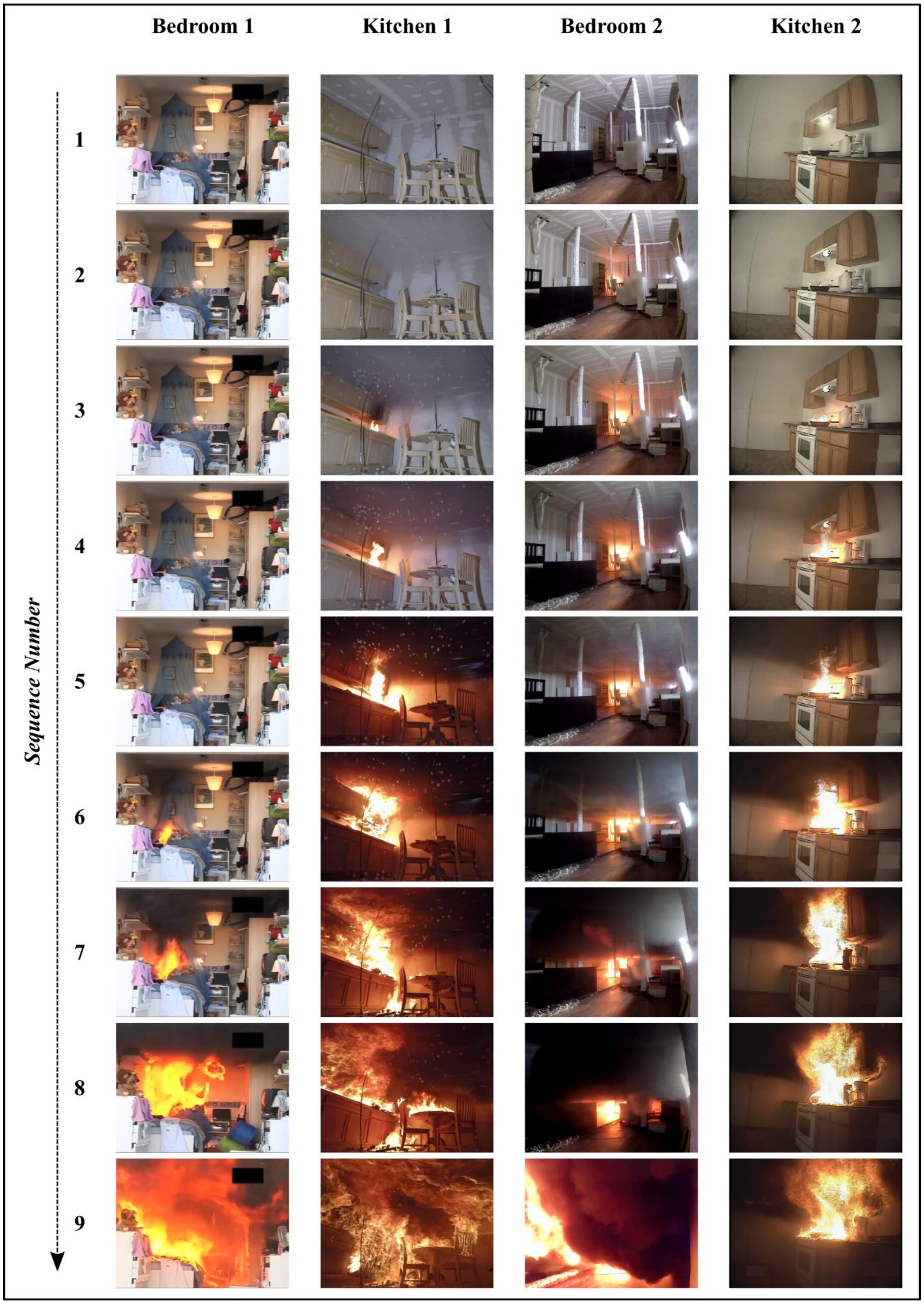


Figure 1. Room fire images presented during the judgment task. The images were selected to display no fire or smoke (#1), low intensity fire, up to high intensity fires (#9). Immediately after each image, a word was presented and participants judged whether the word applied to the image.

Demographic Questionnaire

A series of questions asked participants to provide select demographic information (e.g., age, sex, race). Participants were asked to select from the following options for each demographic factor: biological sex (‘male’, ‘female’), race (‘Asian’, ‘black’, ‘Native American’, ‘white’), and ethnicity (‘Hispanic or Latino’, ‘Not Hispanic or Latino’).

Adult Temperament Questionnaire – Short Form (ATQ-S)

This 77-item questionnaire has been used in previous research to reliably assess different aspects of adult temperament [26]. For each item, individuals judged the extent to which a statement described themselves using a 7-point Likert-type scale (1 = extremely untrue of you; 7 = extremely true of you; e.g., “Sometimes minor events cause me to feel intense happiness.”). If an individual judged a statement to be not relevant, they could respond “not applicable” (this was replaced by the sample mean for that item, following previous research [26]). The responses to the adult temperament questionnaire (ATQ) were compiled into four factor scores, each with subfactors; higher scores indicated higher levels of the factor. Of interest to the present study were the following subfactors: fear (negative affect from anticipated negative event), discomfort (negative affect from sensory stimulation), attentional control (ability to focus and shift attention), and neutral perceptual sensitivity (ability to detect low intensity stimuli) [27].

Risk-Taking Questionnaire 18 Items (RT-18)

Composed of 18 items, the RT-18 has been observed to reliably assess the extent to which young adults engage in risky behaviors [25]. For each item, participants indicated whether they agreed or disagreed that a statement applied to themselves. Scores were summed into two subscales, behavior (extent to which individuals engage in risk-taking behavior; e.g., “I sometimes do “crazy” things just for fun.”) and assessment (extent to which individuals engage in behavior without consideration; e.g., “I often do things on impulse.”), with higher scores indicating higher risk-taking tendencies (scores ranging from 0 to 9 for each subscale).

RESULTS

Task Responses

Prior to analysis, responses for normalcy words were reverse coded to reflect deviation from normalcy, in alignment with responses for risk and protective actions words (resulting responses: normalcy, no risk, no protective action = 0; deviation from normalcy, risk, protective action = 1; $\alpha = 0.05$, two-tailed for all analyses). To assess the extent to which responses to room fires changed as fire size increased and varied with the type of word being applied to an image, a logistic regression was used to model the transition of responses from ‘does not match’ (0) to ‘does match’ (1). Using the ‘glmer’ R package, a generalized liner mixed model (‘logit’ link; ‘bobyqa’ optimizer) with fixed effects of sequence number, word type, and an interaction between sequence number and word type was fitted to the response data with random intercept factors of participant, specific word, and scene, and a random slope and intercept for sequence number by scene.

Using an analysis of deviance, significant main effects of sequence number, $\chi^2(1) = 70.936$, $p < 0.001$, word type, $\chi^2(2) = 70.947$, $p < 0.001$, and interaction effect between sequence number and word category, $\chi^2(2) = 29.871$, $p < 0.001$, were observed. The significant main effects suggested that word prompts were more likely to be judged to apply with later sequence images (i.e., as the fire developed) and, overall, deviation from normalcy prompts were more likely to be judged to apply than ‘risk’ or protective action prompts. The significant interaction between the word type and the sequence suggested that the rate of change of the applicability of that word with respect to sequence number varied by word type. Linear contrast codes used to examine the interaction between sequence number and word category were based on the PADM model, namely that the rate of change for deviation from normalcy (contrast code: 1) would be greater than identifying risk (contrast code: 0), followed by taking protective action (contrast code: -1; analysis used 'glht' R-package). Post-hoc tests revealed a significant effect for the planned contrasts, $z = 4.204$, $p < 0.001$: the largest rate of change was observed for deviation from normalcy, a smaller rate of change for risk, and the smallest rate of change for protective action (see Figure 2).

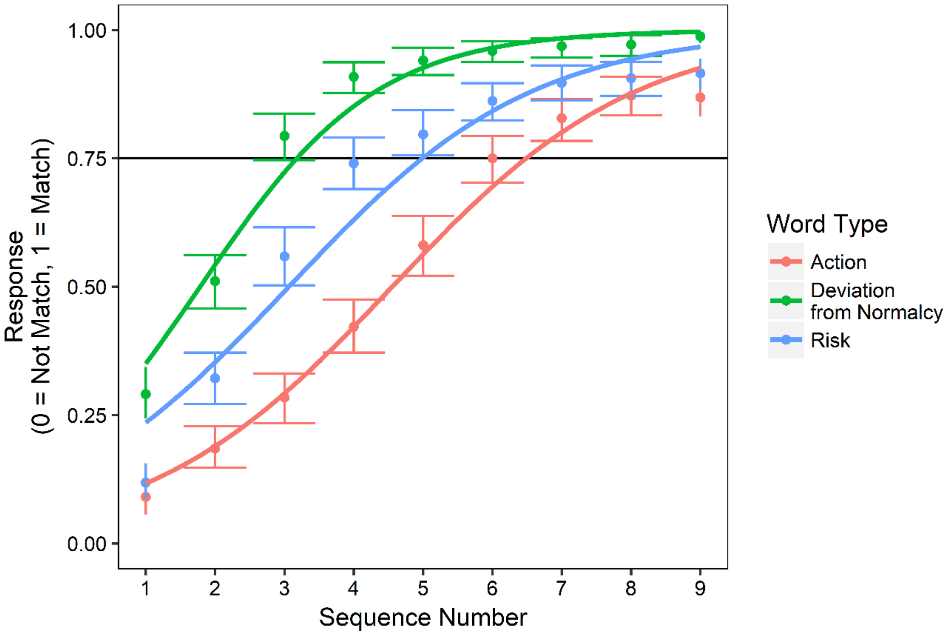


Figure 2. Response patterns on judgment task across word types. Data points reflect mean performance for each image sequence number (95% confidence interval error bars). Solid lines reflect logistic curves fitted to the data points for each word type. The horizontal line reflects 75% performance.

Threshold and Questionnaire Scores

To examine individual differences in performance on the judgment task, models were used to estimate psychophysical thresholds for each word type. Specifically, for each participant, the thresholds estimated the sequence number at which the participant reliably responded (75%) that a word matched subsequent images [28]. To do so, for each participant a logistic regression with sequence number as the predictor and response as the dependent variable was fit for each word type. Due to a poor fit, six participants were dropped from further analysis. Similar to analyses of task responses, planned contrasts using a linear mixed model (random intercept for participant) revealed a linear increase in threshold for word types, $t(58.948) = 7.400$, $p < 0.001$, with participants waiting later in the sequence to reliably judge images as indicating requiring protective action compared to risk, while waiting less for deviations from normalcy compared to risk. Correlational analyses using the Pearson r statistic examined whether relations were present between individual variations in word type thresholds and scores on the ATQ and RT-18 (see **Error! Reference source not found.**). Significant positive correlations were observed between deviation from normalcy and risk thresholds and between risk and protective action thresholds (trend for positive correlation between deviation from normalcy and protective action thresholds). A significant positive correlation was observed between the temperament factor ‘discomfort’ and risk threshold. Although not statistically significant, moderate correlation coefficients were observed between a subset of additional temperament factors and thresholds.

Table 1. Descriptive statistics and correlations between estimated thresholds and dispositional traits (N = 34).

Factor	***Mean***	***SD***	***Median***	***Min***	***Max***	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1-Deviation from normalcy Threshold	2.755	1.05	2.463	1.625	6.099	---	---	---	---	---	---	---	---
2-Risk Threshold	3.378	1.119	3.488	1.567	6.081	0.502**	---	---	---	---	---	---	---
3-Action Threshold	4.530	1.512	4.623	1.87	6.699	0.289†	0.545***	---	---	---	---	---	---
4-ATQ Fear	4.005	0.911	4.143	2.143	5.571	0.075	0.199	0.05	---	---	---	---	---
5- ATQ Discomfort	4.181	1.045	4	2.5	6	0.248	0.381*	0.178	0.446**	---	---	---	---
6-ATQ Attentional Control	4.253	1.026	4.2	1	7	0.236	0.095	-0.055	-0.347*	-0.158	---	---	---
7-ATQ Neutral Perceptual Sensitivity	5.247	0.896	5.3	3.2	6.8	0.238	0.036	0.042	-0.093	-0.021	0.643***	---	---
8-RT-18 Assessment	6.029	2.249	7	1	9	0.036	0.112	-0.001	-0.108	0.288†	0.309†	0.342*	---
9-RT-18 Behavior	3.412	2.439	3	0	9	-0.048	0.012	0.075	0.145	0.188	-0.04	-0.14	0.064

† *p < 0.1*, * *p < 0.05*, ** *p < 0.01*, *** *p < 0.001*

DISCUSSION

In the present study, the extent to which judgments of developing room fires were influenced by situational and dispositional factors was assessed. Performance on the judgment task varied with situational factors, the apparent size of the room fire and word type (i.e., words indicating either, normalcy, risk, or protective action). This pattern of performance aligns with the PADM, suggesting that participants viewed the room fire scenes as indicative of emergencies. Across all word types, as the intensity of the fire cues increased, participants were more likely to judge that words applied to the image. However, the point during the developing room fire at which participants reliably judged that a word matched an image varied by word type. Both logistic model analyses and threshold estimates indicated that participants initially detected fire cues as indicative of a deviation from normalcy (between the second and third images in sequences), then indicative of risk (between the third and fourth images), and finally judged that fire cues indicated required protective action (between fourth and fifth images). This pattern of performance generally aligns with earlier and later stages of the PADM [17] as well as previous observational and self-report evidence collected from fire-related emergencies [20].

Overall, the observed impact of situational factors on judgments of room fires suggests that judgments of presented visual fire cues can be used to examine how humans perceive a fire-related emergency.

When examining performance on the judgment task, participants varied in the point at which they viewed the developing room fires as indicative of risk and requiring protective action. Specifically, individual differences were present in the estimated thresholds for deviation from normalcy, risk, and action, which reflected the point across sequences that participants reliably viewed images as matching the presented words. This provides further evidence that individuals vary in their perceptions and judgments of fire-related cues. When examining connections between word type thresholds, the strongest correlations were between risk and protective action, and deviation from normalcy and risk. These connections align with previous applications of the PADM to fire-related emergencies in two ways. First, the strong positive correlation between risk and protective action thresholds indicates that participants who judged images with less intense fire cues as indicative of risk also judged images as indicative of protective action with less intense cues (and vice versa). This aligns with the stages of the PADM which suggest that when individuals decide a situation poses an imminent risk, they are more likely to take protective action [17]. The correlation between deviation from normalcy and risk thresholds supports previous research indicating that normalcy biases can influence the point at which individuals identify cues as indicative of an emergency [21]; individuals who were later to judge a deviation from normalcy required more intense fire cues before judging the cues as indicative of a risk. Although this aligns with the PADM, a key aspect of the model is the extent to which individuals perceive an emergency as posing a risk to themselves and others [22]. The extent to which participants in the present study simulated themselves, or others, within the images of the room fires may have influenced whether they perceived the fire as posing a risk to themselves or others. This could account for some of the individual differences in thresholds observed in the present study.

When examining individual variations in dispositional traits and judgments of room fires, weak connections were observed. In the present study, temperament and risk-taking measures were selected since previous research has observed that individual differences in these self-reported measures have been connected to real-word behaviors. Of the measures included, the temperament factor discomfort was significantly correlated with variations in risk threshold. Specifically, individuals who had higher discomfort scores took longer to identify images as indicative of risk. With higher levels of discomfort reflecting greater negativity towards sensory stimulation (e.g., greater negative reactions to more intense visual stimulation [27]), this relation suggests that individual differences in dispositional traits may affect perceptions of risk cues in the environment. The lack of statistically significant effects between judgments of room fires and other dispositional traits included in the present study raises questions about the extent to which the present study was able to detect such relations. Specifically, the extent to which participants viewed the images of a fire as posing a risk to other persons may have moderated connections with dispositional traits. Future research that increases the perception that the room fires pose an imminent risk to humans may be better suited to determine whether temperament of individuals influences judgments of fire emergencies.

Future Considerations and Directions

The results of the present study provide initial evidence that judgments of still images of developing room fires align with past observations of human responses to emergencies. Although the words used during judgments for the present study were selected to reflect earlier versus later decision making processes present in the PADM, it remains to be determined which stages or phases of the model the words most directly align with. Earlier phases of the PADM focus on the perception of cues in the environment that are indicative of an emergency with later phases emphasizing the interpretation of the cues [22]. Although the sequence in which words were judged to apply to developing fires generally aligned the order of PADM phases, questions remain regarding the extent to which participants interpreted the words selected to reflect risk and protective action as indicative of a threat to others. By design, the fire cues included in the present study were limited to visual forms and did not include human occupants. With evidence that decision making during emergencies can be influenced by the perceived risk of the hazard to the self and others, this may have influenced the judgments of participants. Future research that utilizes more realistic or immersive cues and/or information sources (e.g., virtual reality) and more descriptive emergency scenarios may be able to increase the perceived risk to the self and others. For example, with evidence that using virtual reality can increase perceived presence in an environment [29], presenting a three-dimensional virtual room fire, with the participant situated inside the room, may be able to increase perceived presence and sense of risk to the individual. Alternatively, providing participants with scenarios that emphasize the potential risk to others (e.g., describing a bedroom room fire as occurring adjacent to another bedroom where a child is currently sleeping), could lead to a different pattern of responses and relations to dispositional traits. Additionally, the stimuli in the present study were strictly within the visual modality. When conducting interviews with survivors of fire emergencies, many report that cues within other sensory modalities, such as smell, were noticed in addition to visual cues [8]. Open questions remain as to whether patterns in performance observed in the present study would be observed when using fire cues of different, or multiple, sensory modalities.

The stimuli selected for the present study focused on bedroom and kitchen fires that were video recorded as part of controlled burns. Although the videos were of real fires, some aspects of the videos may have led participants to view the rooms as atypical. For example, the angle of the video camera (e.g., near floor) and inclusion of measurement devices and recording instruments in a subset of the videos may have made it difficult for participants to visualize themselves as present in the room. Future research should include additional room fire scenes that are likely to be familiar with the targeted population of participants. Furthermore, there were differences across the room scenes with regard to the visual extent of the fires and sequence number. To account for differences in the rate at which visual cues of fire intensity changed across room fire scenes, mixed models that included a random effect of scene were utilized. Although this statistical procedure was used to minimize the effect of cross-scene differences, future research should more systematically match visual cues of fire intensity across different rooms and environments. Furthermore, the words included in the present study were selected to reflect different stages of the PADM, but future research could include a range of words, or phrases, to examine the impact on judgment performance.

Overall, the present study provides a decision-making task for examining how individuals judge fire cues in different situations. With evidence that judgment performance aligned with the PADM, future research can make use of the paradigm to continue to examine different facets of human perception and processing of cues present during fire, and non-fire, emergencies.

References

[1] G. V. Hadjisophocleous, N. Benichou, A.S. Tamim, Literature review of performance-based fire codes and design environment, J. Fire Prot. Eng. 9 (1998) 12–40. doi:10.1177/104239159800900102.

[2] B.J. Meacham, R.L.P. Custer, Performance-based fire safety engineering: An introduction of basic concepts, J. Fire Prot. Eng. 7 (2007) 35–54. doi:10.1177/104239159500700201.

[3] S. Gwynne, E.R. Rosenbaum, Employing the hydraulic model in assessing emergency movement, in: M. Hurley (Ed.), SFPE Handb. Fire Prot. Eng. Fifth Ed., 2016: pp. 2115–2151. doi:10.1007/978-1-4939-2565-0_59.

[4] P. Reneke, Evacuation Decision Model (NIST IR 7914), Gaithersburg, MD, 2013.

[5] R. Lovreglio, E. Ronchi, D. Nilsson, An Evacuation Decision Model based on perceived risk, social influence and behavioural uncertainty, Simul. Model. Pract. Theory. 66 (2016) 226–242. doi:10.1016/j.simpat.2016.03.006.

[6] R. Lovreglio, E. Ronchi, D. Nilsson, A model of the decision-making process during pre-evacuation, Fire Saf. J. 78 (2015) 168–179. doi:10.1016/j.firesaf.2015.07.001.

[7] J. Bryan, A selected historical review of human behavior in fire, Fire Prot. Eng. 16 (2002) 4–10.

[8] M. Kobes, I. Helsloot, B. de Vries, J.G. Post, Building safety and human behaviour in fire: A literature review, Fire Saf. J. 45 (2010) 1–11. doi:10.1016/j.firesaf.2009.08.005.

[9] P. Brennan, Timing Human Response In Real Fires, in: Proc. Fifth Int. Symp. Fire Saf. Sci. (IAFSS 5), International Association for Fire Safety Science, 1997: pp. 807–818. doi:10.3801/iafss.fss.5-807.

[10] P. Brennan, Modelling cue recognition and pre-evacuation response, in: Proc. Sixth Int. Symp. Fire Saf. Sci. (IAFSS 6), International Association for Fire Safety Science, 2000: pp. 1029–1040. doi:10.3801/IAFSS.FSS.6-1029.

[11] M. Liu, S.M. Lo, The quantitative investigation on people’s pre-evacuation behavior under fire, Autom. Constr. 20 (2011) 620–628. doi:10.1016/j.autcon.2010.12.004.

[12] E. Kuligowski, Burning down the silos: Integrating new perspectives from the social sciences into human behavior in fire research, Fire Mater. 41 (2017) 389–411. doi:10.1002/fam.2392.

[13] P.G. Wood, The Behaviour of People in Fires, 1972.

[14] J. Bryan, Smoke as a determinant of human behavior in fire situations, University of Maryland, College Park, 1977.

[15] O.F. Thompson, E.R. Galea, L.M. Hulse, A review of the literature on human behaviour in dwelling fires, Saf. Sci. 109 (2018) 303–312. doi:10.1016/j.ssci.2018.06.016.

[16] E.D. Kuligowski, S.M.V. Gwynne, M.J. Kinsey, L. Hulse, Guidance for the model user on representing human behavior in egress models, Fire Technol. 53 (2017) 649–672. doi:10.1007/s10694-016-0586-2.

[17] E. Kuligowski, Predicting human behavior during fires, Fire Technol. 49 (2013) 101–120. doi:10.1007/s10694-011-0245-6.

[18] M.K. Lindell, R.W. Perry, The protective action decision model: Theoretical modifications and additional evidence, Risk Anal. 32 (2012) 616–632. doi:10.1111/j.1539-6924.2011.01647.x.

[19] N.C. McConnell, K.E. Boyce, J. Shields, E.R. Galea, R.C. Day, L.M. Hulse, The UK 9/11 evacuation study: Analysis of survivors’ recognition and response phase in WTC1, Fire Saf. J. 45 (2010) 21–34. doi:10.1016/j.firesaf.2009.09.001.

[20] M.T. Kinateder, E.D. Kuligowski, P.A. Reneke, R.D. Peacock, Risk perception in fire evacuation behavior revisited: definitions, related concepts, and empirical evidence, Fire Sci. Rev. 4 (2015) 1–26. doi:10.1186/s40038-014-0005-z.

[21] M.J. Kinsey, S.M.V. Gwynne, E.D. Kuligowski, M. Kinateder, Cognitive Biases Within Decision Making During Fire Evacuations, Fire Technol. (2018) 1–22. doi:10.1007/s10694-018-0708-0.

[22] E.D. Kuligowski, The Process of Human Behavior in Fires, 2009.

[23] D.C. Schwebel, K.K. Ball, J. Severson, B.K. Barton, M. Rizzo, S.M. Viamonte, Individual difference factors in risky driving among older adults, J. Safety Res. 38 (2007) 501–509. doi:10.1016/j.jsr.2007.04.005.

[24] D.C. Schwebel, D. Stavrinos, E.M. Kongable, Attentional control, high intensity pleasure, and risky pedestrian behavior in college students, Accid. Anal. Prev. 41 (2009) 658–661. doi:10.1016/j.aap.2009.03.003.

[25] L. de Haan, E. Kuipers, Y. Kuerten, M. van Laar, B. Olivier, J.C. Verster, The RT-18: a new screening tool to assess young adult risk-taking behavior., Int. J. Gen. Med. 4 (2011) 575–84. doi:10.2147/IJGM.S23603.

[26] M.K. Rothbart, D.E. Evans, S.A. Ahadi, Temperament and personality: Origins and outcomes, J. Pers. Soc. Psychol. (2000). doi:10.1037/0022-3514.78.1.122.

[27] D. Derryberry, M.K. Rothbart, Arousal, affect, and attention as components of temperament., J. Pers. Soc. Psychol. 55 (1988) 958–966. doi:10.1037/0022-3514.55.6.958.

[28] K. Knoblauch, L.T. Maloney, Modeling psychophysical data in R, Springer-Verlag, New York, NY, 2012. doi:10.1007/978-1-4614-4475-6.

[29] D. Vastfjall, The subjective sense of presence, emotion recognition, and experienced emotions in auditory virtual environments, CyberPsychology Behav. 6 (2003) 181–188.

A Simulation Platform to Study the Human Body Communication Channel

Katjana Krhac[1], Kamran Sayrafian[2], Gregory Noetscher[3], Dina Simunic[1]

[1]University of Zagreb
Zagreb, Croatia

[2]National Institute of Standards &
Technology, Gaithersburg, MD, USA

[3]NEVA Electromagnetics, LLC
Yarmouth Port, MA, USA

Abstract— Human Body Communication (HBC) is an attractive low complexity technology with promising applications in wearable biomedical sensors. In this paper, a simple parametric model based on the finite-element method (FEM) using a full human body model is developed to virtually emulate and examine the HBC channel. FEM allows better modeling and quantification of the underlying physical phenomena including the impact of the human body for the desired applications. By adjusting the parameters of the model, a good match with the limited measurement results in the literature is observed. Having a flexible and customizable simulation platform could be very helpful to better understand the communication medium for capacitively coupled electrodes in HBC. This knowledge, in turn, leads to better transceiver design for given applications. The platform presented here can also be extended to study communication channel characteristics when the HBC mechanism is used by an implant device.

Keywords- Human body communication, Computational human body models, Capacitive coupling

I. Introduction

Human Body Communication (HBC) is one of the wireless technologies defined by the IEEE 802.15.6 standard on body area networking [1]. In HBC, the human body is used as a communication medium between a pair of transmitter and receiver electrodes that are placed on the body surface. The low complexity and energy consumption are among the reasons that make this technology attractive for wearable and implantable devices. Also, as the transmitted data is mostly confined to the human body area, there is less chance of unauthorized access; and therefore, better security is expected compared to other wireless technologies used for body area networks. In the literature, the general technology has also been referred to as Body Channel Communications (BCC) or Intra-Body Communications (IBC). These non-RF communication mechanisms include capacitive (or equivalently electric field) and galvanic signal coupling.

In the capacitive coupling method, the electrical signal that is applied to the human body (i.e. forward path) is capacitively coupled through the air or the environment where the body is located (i.e. return path). Alternatively, in the galvanic coupling method, the human body would act as a waveguide for the signal that is injected by the alternating current into the body. The term HBC, as outlined in the IEEE802.15.6, mainly refers to the capacitive coupling methodology. The underlying concept behind HBC is the fact that in the presence of a weak electric field, the human body can act as a signal guide to capacitively couple two electrodes that are in contact with the body surface. The coupling through the body is achieved with much less attenuation compared to the free space. There have been several studies by researchers to better understand and characterize the HBC channel in the past 10 years. These studies are mostly in the form of measurement campaigns considering different scenarios. As the environment around the human body (and possibly the body posture) directly affects the wireless link between two HBC electrodes, some discrepancies are often observed among the physical measurement results reported in the literature. Also, due to the variables and parameters that could impact the return path, developing a comprehensive simulation platform that can adequately model the HBC channel is quite challenging. Methodologies that have been used to investigate the electric field propagation mechanism and modeling the underlying body channel include RC circuits, Finite-Element Method (FEM), Circuit-coupled FEM as well as FDTD (Finite Difference Time Domain) method to model the communication link [3 - 8].

Circuit models representing the HBC channel have been developed in [5, 6]. The reported results for the channel gain (or equivalently pathloss) in the range of 1 to 100 MHz show a bandpass profile with a peak somewhere in the 40-60 MHz interval. In [7], the authors have developed a circuit-coupled finite-element method (FEM) model of the HBC channel. A multi-layer FEM model was used to represent the human forearm. The parasitic effects of the printed circuit board (PCB) and the return path were modeled as circuit elements. While measurement results for the channel attenuation versus frequency still showed a bandpass profile, simulation results show monotonically increasing with a much milder bandpass profile for various parameters values. In [8], the authors concluded that measurements of the pathloss in HBC systems strongly depend on the instrumentation configurations. Using a small battery-powered transmitter and receiver to measure pathloss, they showed that inclusion of any additional ground plane, even isolated by baluns, serves to underestimate the resulting path loss by up to 33.6 dB. Their pathloss results again showed a bandpass profile with a peak around 70-75 MHz.

The common element in most of the results published in the past several years is the passband profile shape of the channel attenuation in HBC within the range 1-100 MHz. As pointed out earlier, depending on the instrumentation or methodology that was used for channel measurement, the

Sayrafian, Kamran; Krhac, Katjana; Noetscher, Gregory; Simunic, Dina. "A Simulation Platform to Study the Human Body Communication Channel." Paper presented at IEEE Engineering in Medicine and Biology Conference (EMBC 2019), Berlin, DE. July 23, 2019 - July 26, 2019.

location of the peak frequency in the passband profile varies from 40 to 70 MHz. In addition, discrepancies are also observed on the average magnitude of the forward transmission coefficient. This is due to the specifics of the model used for the parasitic return path in the HBC system. Our objective in this research was to develop a simple parametric FEM-based model that 1) can capture the fundamental concept of HBC operation/channel; 2) can be adjusted to emulate a specific measurement scenario; and 3) can be easily extended to study implant HBC channel.

The rest of the paper is organized as follows. Section II describes a computational 3D human body model that is used for this study along with a simple model of the HBC electrodes and communication links. Section III provides simulation results for several on-body scenarios and a brief discussion on comparison with existing physical measurements in the literature. Finally, conclusions and our plans for future work are described in section IV.

II. Simulation platform

Modeling of a HBC channel is a challenging task due to the many parameters that can possibly affect the characteristics of the communication link. Among those, we can point to size and shape of the electrodes, locations and the distance between the RX/TX electrodes, separation between the signal and ground plates of each electrode, body posture, dielectric properties of the human tissues that are in contact with the electrodes and possibly the environment surrounding the human body.

A novel 3D immersive platform to study wireless channels in body area networks has been developed at the Information Technology Laboratory of the National Institute of Standards & Technology (NIST) [2]. A 3D computational human body model is one of the main components of this platform that can be customized for the desired application. The body model includes frequency dependent dielectric properties of 300+ parts in a male human body. These properties are also user-definable if specific changes or modifications are desired. The human body model has a resolution of 2 mm. To study HBC, this computational model has been augmented with a skin shell that fits over the exterior body mesh. In addition, variable fat shells (reflecting skinny, average and obese persons) have been added to the model. This will allow us to study the potential impact of the fat layer on the forward path attenuation of a HBC channel.

The electrodes were modeled as two metal plates; one in contact with the skin and the other located directly above and floating in the air. To ensure the full contact of electrodes with the skin tissue in the body model, a small patch (i.e. brick) of skin material has been added directly underneath the electrode to unite the signal plate with the skin exterior. This ensures that the surface of the electrode and the skin are fully coincident. Size, plate separation, distance between the receiver/transmitter electrodes are design variables; and therefore, the impact of each one of these parameters can be easily investigated in our model. Figure 1 shows an example of two electrodes placed on the right arm of the human body model in our platform.

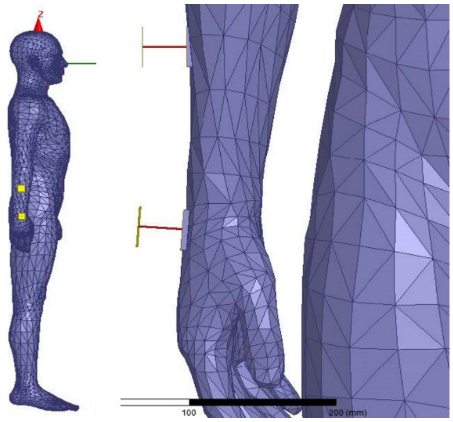


Figure 1. HBC electrodes on the right arm

The biggest complexity in modeling the HBC channel is incorporating the impact of the parasitic return path and characteristic impedances of the electronic circuits generating the signal (or in case of physical measurement, parasitic of the probe PCB). In our FEM-based model, the coupling of the electrodes through the air (i.e. return path) is modeled by a capacitor (C_{Ret}). This is similar to the circuit or circuit-enabled models proposed in [7]. However, we have implemented the capacitive return path using RLC boundaries methodology in ANSYS. The signal leakage path between the electrode plates has also been modeled with capacitor C_L. The characteristic impedance of the source has been modeled by a cascade of resistor R_T and inductor L_T. These elements are schematically shown in Figure 2.

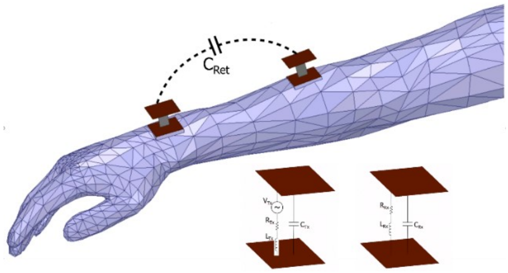


Figure 2. System schematic

Using ANSYS HFSS[1] electromagnetic solver, a variety of different quantities such as the magnitudes of the electric field (inside, on the surface and outside of the body) and the Scattering parameters (e.g. S_{21}) between the two electrodes can be calculated. For example, Figure 3 highlights the basic principle of HBC by displaying the electric field distribution. Fig. 3(a), on the left, shows the distribution of the magnitude of the electric field when the transmitting electrode is not in contact with the human body (i.e. operating in the air). On the other hand, when the electrode is place on the human arm (as

[1] HFSS is a product of ANSYS, Inc. HFSS has been used in this research to foster research and understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that this product is necessarily the best available for the purpose.

seen in Fig. 3(b)), the electric field extends over the entire body surface. This ensures much higher received signal strength; and therefore, better communication channel between the two electrodes. A signal frequency of 50 MHz has been used for the result shown in Figure 3.

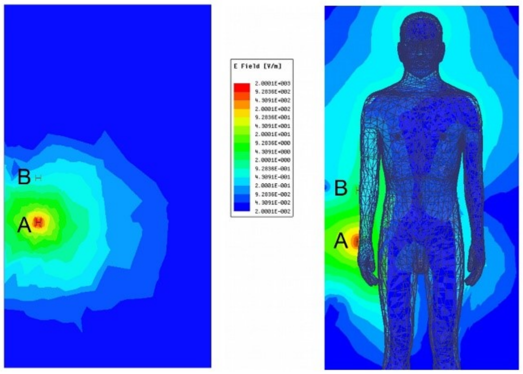


Figure 3. Electric field distribution around the transmitter electrode (a) without the human body (b) with the human body

The HBC platform mentioned above enables us to run different parametric studies to find the matching values of the parasitic elements as well as investigate the impact of electrodes size, location, and body placement for various applications. In the next section, we point out to some of these results.

III. Results

The frequency range of interest in HBC is typically 1-100 MHz. Higher frequencies could result in the human body acting as an antenna and are also susceptible to external radiation [9]. Therefore, for frequencies higher than 100 MHz, significant channel variation and lower efficiency of the communication system can be expected. Using the simulation platform discussed in the previous section, the HBC channel attenuation can be measured by the forward transmission coefficient S_{21} for various scenarios. As mentioned earlier, limited measurement results in the literature indicate a bandpass profile with varying location of the peak frequency. The bandpass profile is mostly influenced by the capacitive leakage path as well as the parasitic inductance L_{Tx}.

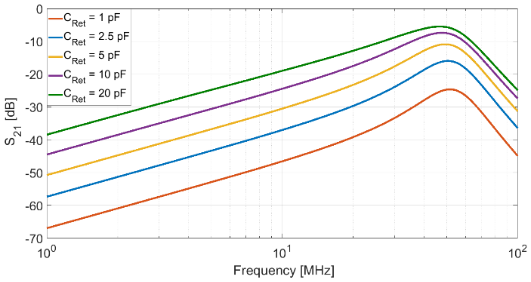


Figure 4. Forward transmission coefficient for various C_{Ret}

To show the impact of these parameters, consider the scenario depicted in Figure 1 with a distance of 15 cm between the two electrodes on the arm. Figure 4 displays the simulation results for various values of the capacitive return path (C_{Ret}) when L_{Tx}=225 nH. Similarly, Figure 5 reflects the impact of varying L_{Tx} while C_{Ret} is fixed. As observed, the bandpass profile shape and values for the S_{21} can be tuned and optimized to match the physical measurement results published in the literature. In fact, for 15 cm separation between the electrodes, C_{Ret}=5 pF, L_{Tx}=220 nH and C_L=35 pF, the simulation results match well with the reported physical measurement results in [5, 6]. The source and load impedances were assumed to be 50 Ohms. We also investigated the S_{21} results for distances of 30 cm and 45 cm as the receiver electrode was moved further up the arm and closer to the shoulder. The same values for the parasitic capacitance and inductance resulted in the best match with the reported measurements in [3, 4].

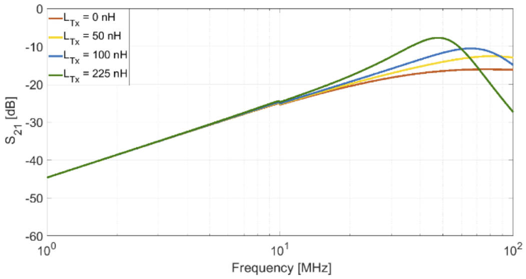


Figure 5. Forward transmission coefficient for various parasitic inductance

Looking at the results for electrode separations of 15, 30 and 45 cm, a slight degradation of the S_{21} versus distance is observed (see Fig. 6). However, the parasitic capacitance representing the return path seems to be the same. The S_{21} degradation is due to the higher attenuation of the forward path, i.e. the path through the body. For the three scenarios considered so far, the forward and return path distances are approximately the same as the receiver electrode is moved away from the transmitter along the straight human arm. It is worth noting that there could be scenarios where the forward and return path distances are quite different, for example, when the electrodes are located on the left and right wrists, and the arms are held straight down along the body (or even close to each other in front of the body).

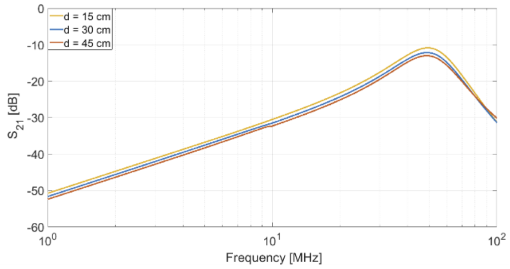


Figure 6. Forward transmission coefficient for various electrodes distances

To the best of our knowledge, there are no results in the literature that highlights the impact of this discrepancy between forward and return path distances. Our conjecture is that the parasitic capacitance could be a function of the return path distance. For example, if the electrodes are located on the separate wrists, the magnitude of S_{21} will change as the person under experiment move his hands closer or farther away from each other. Also, another ambiguity regarding the forward path distance is whether this distance is considered as the shortest path through the human body (i.e. in-body distance) or on the body surface. For example, consider two HBC electrodes, one located on the chest area while the other is on the back side of the body. Depending on how the electric field is distributed, the length of the forward path could be the straight line through the body connecting the electrodes or the distance around the body surface separating the two electrodes. To better understand this issue, we first need to eliminate the impact of the return path which can be accomplished by short-circuiting the ground plates of the electrodes with a perfect conductor. This was first done for the scenario in Figure 1 and the resulting S_{21} versus frequency (which is now representing only the attenuation through the forward path) is shown in Figure 7.

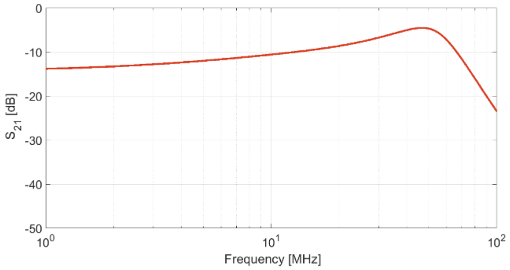


Figure 7. Forward path attenuation for the scenario shown in Fig. 1

As observed, the body seems to be a uniform attenuator (for frequencies below 100 MHz) with approximately 10-15 dB loss when the distance between TX and RX electrodes is about 15 cm. This level of attenuation was also observed through physical measurement in [7].

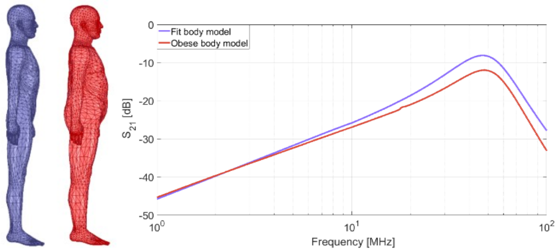


Figure 8. Forward transmission coefficient for various body shapes

The material properties of the body tissues that are along the forward path of the HBC channel can also impact the signal attenuation. To observe this, we conducted a simulation using two body models representing a fit versus obese person (see Fig. 8). The electrodes were placed on the stomach area where a heavier concentration of fat exists. The electrode separation was chosen to be 20 cm. The S_{21} for the two human body models are shown in Figure 8. The result indicates higher attenuation for the obese person, which can be explained by the lower conductivity of the fat tissue versus muscle for the frequency range of 10 to 100 MHz.

IV. Conclusions and Future Work

A FEM-based parametric simulation platform including a full 3D computational human body model has been presented in this paper. The S_{21} results match the bandpass HBC channel attenuation profile reported in the literature with appropriate adjustment of the model parameters. The platform allows researchers to further study the HBC channel by considering variable electrode size and plate separation, placement on the body, as well as designing virtual experiments to better understand the impact of variable return path distance for a fixed forward path through the human body. The authors also intend to continue this research by further extending the computational human body model to include various postures. Further studies on the distribution of the electric field inside the body would also be necessary to investigate the potential impact of using HBC on implant devices such as pacemakers.

Also, as mentioned earlier, low complexity and energy consumption of HBC will also make this technology an attractive alternative for implantable devices. Since implants are completely located inside the human body, physical measurements are no longer possible to examine the channel. Therefore, a simulation platform including a full human body model will be very useful to study and characterize the implant HBC channel. The implant communication link is less affected by the environmental variables as both forward and return paths are confined within the human body. Development of the platform discussed in this paper is the first step toward a comprehensive study of the implant HBC channel.

References

[1] IEEE Standard for Local and Metropolitan Area Networks, IEEE 802.15.6-2012 – Part 15.6: Wireless Body Area networks, 2012

[2] K. Sayrafian, J. Hagedorn, W. B. Yang, J Terrill, "A Virtual Reality Platform to Study RF Propagation in Body Area Networks", IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), Kosice, Slovakia, Dec. 2-5, 2012

[3] M. A. Callejon, D. Naranjo-Hernandez, J. Reina-Tosina, and L.M. Roa, "A comprehensive study into intrabody communication measurements," IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 9, pp. 2446–2455, 2013

[4] Ž. Lučev Vasić, "Intrabody communication based on capacitive method" Doctoral thesis, University of Zagreb, Faculty of Electrical Engineering and Computing, 2014

[5] M. Amparo Callej´on, D. Naranjo-Hern´andez, J. Reina-Tosina, L. M. Roa, "Distributed circuit modeling of galvanic and capacitive coupling for intrabody communication," IEEE Transactions on Biomedical Engineering, vol. 59, no. 12, pp.3263–3269, 2012

[6] M. Pereira, G. Alvarez-Botero, F. Rangel de Sousa, "Characterization and Modeling of the Capacitive HBC Channel," IEEE Transactions on Instrumentation & Measurement, Vol. 64, no.10, Oct. 2015

[7] R. Xu, H. Zhu, and J. Yuan, "Electric-field intrabody communication channel modeling with finite-element method," IEEE Transactions on Biomedical Engineering, vol. 58, no. 3, pp. 705–712, 2011

[8] J. Park, H. Garudadri, and P. P. Mercier, "Channel Modeling of Miniaturized Battery-Powered Capacitive Human Body Communication Systems," IEEE Transactions on Biomedical Engineering, vol. 64, no. 2, pp. 452–462, 2017

[9] B. Kibret, A. K. Teshome, D. T. H. Lai, "Human Body as Antenna and Its Effect on Human Body Communications", Progress in Electromagnetics Research, Vol. 148, 2014

Sayrafian, Kamran; Krhac, Katjana; Noetscher, Gregory; Simunic, Dina. "A Simulation Platform to Study the Human Body Communication Channel." Paper presented at IEEE Engineering in Medicine and Biology Conference (EMBC 2019), Berlin, DE. July 23, 2019 - July 26, 2019.

A strategy for handling aberration in Spherical Neutron Polarimetry

Jacob Tosado,[a] **Wangchun Chen**[b c] **and Efrain E. Rodriguez**[a]

[a]Department of Chemistry & Biochemistry, University of Maryland,
College Park, MD 20742.
[b]NIST Center for Neutron Research, National Institute of Standards and Technology,
100 Bureau Drive, Gaithersburg, MD 20899.
[c]Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742.

E-mail: jtosado@umd.edu

Abstract. We present a strategy for identifying and correcting for aberration effects in Spherical Neutron Polarimetry. The transformation of the neutron beam polarization vector due to scattering from a material is determined with Spherical Neutron Polarimetry. This neutron scattering technique measures the three cardinal components of the scattered polarization for any chosen cardinal direction of the incident polarization for a given Bragg reflection. As a consequence, the instrumentation required for this technique is desired to be capable of measuring the three-dimensional polarization vector over the sphere. As with all instrumentation, the field of measurement is subject to aberration which must be characterized.

1. Introduction

Spherical Neutron Polarimetry (SNP) is a neutron scattering technique that measures the nine elements of the polarization property tensor of a material [1]. These nine elements are derived from the three-dimensional neutron beam polarization vector with Cartesian components $P_j = (I_+ - I_-)/(I_+ + I_-)$, for $j = 1$, 2 or 3 corresponding to X, Y or Z. The terms, $I_\pm$, correspond to the measured intensity of the $\pm 1/2$ spin eigenstates of the neutron angular momentum in the j^{th} direction. This measurement technique was first fully realized in 1989 at the Institut Laue-Langevin [2] and has since been revised, [3] expanded upon [4, 5] and developed elsewhere [6]. Figure 1 illustrates in a three step process the general layout of SNP. In the first step, the polarization direction of the incident polarized neutron beam is set to one of the j^{th} Cartesian directions. In the second step, this oriented polarized beam scatters, through Bragg reflection from a crystal, which changes the neutron beam's initial momentum, $\mathbf{k}_1$, to some final momentum, $\mathbf{k}_2$. In the third step, for a given Bragg reflection, the axis of measurement is set to measure one Cartesian component of the scattered neutron beam. Orienting and measuring all possible directional combinations yields the nine component polarization property tensor for that material at a specific Bragg reflection.

The Cartesian axis of measurement and orientation is fixed relative to the scattering vector, $\mathbf{Q}$, which specifies a Bragg reflection and is defined as the change in momentum between the incident and scattered neutron beams (i.e., $\mathbf{Q} = \mathbf{k}_2 - \mathbf{k}_1$). We call this frame of reference the Q-frame. To easily move a neutron detection apparatus about a crystal and thereby measure

Tosado, Jacob; Chen, Wangchun; Rodriguez, Efrain E. "A Strategy for Handling Aberration in Spherical Neutron Polarimetry." Paper presented at 12th International Conference on Polarised Neutrons for Condensed Matter Investigations Journal of Physics: Conference Series, Abingdon, GB. July 03, 2018 - July 06, 2018.

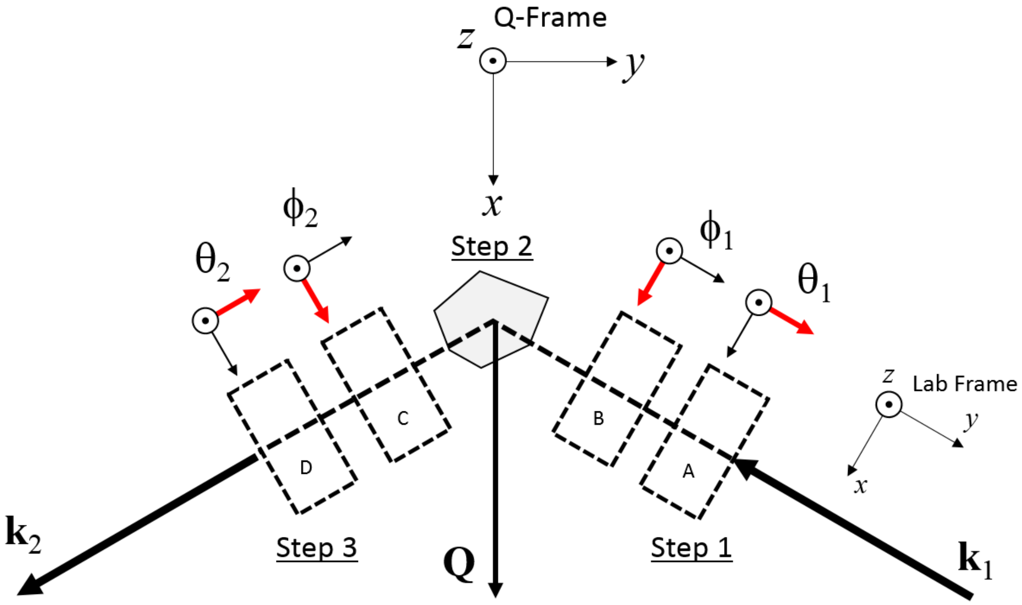


Figure 1. Coordinate system of the SNP setup as seen from the Q-frame. Here, the XY-plane is the scattering plane. The relative placement of polarization control instrumentation is outlined with boxes. Adjacent to each box, the respective local coordinate system and local control parameter are depicted. For each local coordinate system, the axis of rotation for the control parameter is highlighted in red. The letters distinguish the instrumentation as follows: **A)** Local rotation about the y_1 axis, $R_{y_1}(\theta_1)$ **B)** Local rotation about the x_1 axis, $R_{x_1}(\phi_1)$ **C)** Local rotation about the x_2 axis, $R_{x_2}(\phi_2)$ **D)** Local rotation within the XZ-plane, $B_{xz}(\theta_2)$

different **Q**, the possible directions of **Q** are typically restricted to rotations about a single axis fixed in the lab frame. The lab frame may be thought of as being aligned with, $\mathbf{k}_1$, the direction of the incident neutron beam. The resulting plane through which **Q** rotates defines what is called the scattering plane.[1] The Cartesian directions are then typically labeled X, Y and Z where X is parallel to **Q**, Z is vertical and perpendicular to the scattering plane and, Y completes the orthogonal set. Upon scattering, the neutron beam's initial state polarization is transformed by interacting with the magnetic structure of the crystal to some final state polarization. That transformation is defined by the Blume-Maleyev tensor [7].

2. Calibration

The instruments that orient the incident polarization and set the direction of the measurement axis each have local independent coordinate systems. Through calibration, the local coordinate system of each instrument is transformed into the Q-frame. As such, calibration requires the precise mutual alignment of each coordinate system and the precise characterization of distortions inherent within the SNP apparatus. Each SNP component device(s) can be identified with respect to the polarization degree of freedom that it controls and/or measures. As a result,

[1] Throughout this article we generally assume that the scattering plane is level with the floor of the laboratory. This is an important assumption in that it places real restrictions on the overall geometry of an SNP apparatus which naturally influences the potential distortion in the measurement.

there are four main devices corresponding to four degrees of freedom: two degrees for the incident beam and two degrees for the scattered beam, which must be mutually aligned. To understand alignment one can describe the measurement of a single component, P_j, of the polarization vector in terms of the rotation of an initial state polarization vector and its dot product with the measurement axis in the Q-frame. This is given by the relationship,

$$P_j = \mathbf{B}(\Theta_2, \Phi_2) \cdot T \;\; R(\Theta_1, \Phi_1)\mathbf{P}_o \tag{1}$$

Here $\mathbf{P}_o$ is the initial state polarization, T is the Blume-Maleyev tensor, $R(\Theta_1, \Phi_1)$ is the net rotation of the initial state polarization by the instrumentation and $\mathbf{B}(\Theta_2, \Phi_2)$ is the axis of measurement. The spherical coordinates $\Theta_{1,2}$ and $\Phi_{1,2}$ are angles about the Q-frame Y and X axes, respectively. Implicitly, T, Θ_1, Φ_1, Θ_2 and Φ_2 are all dependent on $\mathbf{Q}$. For simplicity we will first consider one value of $\mathbf{Q}$ and later discuss other directions. From Equation 1 the three steps of Figure 1 can be distinguished mathematically as,

Step 1.	$R(\Theta_1, \Phi_1)\mathbf{P}_o$	Directing the incident polarization
Step 2.	T	Scattering from a crystal
Step 3.	$\mathbf{B}(\Theta_2, \Phi_2)$	Measurement of the scattered polarization

The first part in calibration is to remove effects due to magnetic scattering from a crystal so as to isolate the character of the apparatus. This can be achieved by either removing the crystal from the SNP apparatus altogether and measuring the polarization of the purely transmitted beam or by measuring the scattered polarization of a pure nuclear Bragg peak. In either case, having all effects resulting from magnetic scattering properly removed will reduce T to the identity operator.

To understand the calibration of each component instrument we first decompose the rotation operator, $R(\Theta_1, \Phi_1)$, into three operators: $R_{in}(\mathbf{Q})$, $R_{y_1}(\theta_1)$ and $R_{x_1}(\phi_1)$ where, $R(\Theta_1, \Phi_1) = R_{in}(\mathbf{Q})R_{x_1}(\phi_1)R_{y_1}(\theta_1)$. Here, $R_{y_1}(\theta_1)$ and $R_{x_1}(\phi_1)$ represent a rotation about the local y_1 and x_1 axes, respectively, and $R_{in}(\mathbf{Q})$ represents the net transformation into the Q-frame. In Figure 1 the relative orientations of the local axis as seen from the Q-frame are depicted. The measurement axis decomposes likewise into three operators: $\mathbf{B}_{xz}(\theta_2)$, $R_{x_2}(\phi_2)$, and $R_{out}(\mathbf{Q})$ where $\mathbf{B}(\Theta_2, \Phi_2) = \mathbf{B}_{xz}(\theta_2)R_{x_2}(\phi_2)R_{out}(\mathbf{Q})$. Here $\mathbf{B}_{xz}(\theta_2)$ is a local planar measurement field in the XZ-plane, $R_{x_2}(\phi_2)$ is a local rotation of the scattered polarization into that field, and $R_{out}(\mathbf{Q})$ is the net transformation out of the Q-frame. To simplify matters further, we only consider the transmitted beam method for calibration such that T, $R_{in}(\mathbf{Q})$ and $R_{out}(\mathbf{Q})$ each reduce to the identity operator.[2] The measured polarization of the purely transmitted beam is now,

$$P_j = \mathbf{B}_{xz}(\theta_2)R_{x_2}(\phi_2) \cdot R_{x_1}(\phi_1)R_{y_1}(\theta_1)\mathbf{P}_o \tag{2}$$

From this Equation we can define two domains on either the incident polarization or on the measurement of the scattered polarization. The first is the domain of control, which describes all possible orientations of the incident polarization,

$$\mathbf{S^2}_{con} = \{(\theta_1, \phi_1) \;\mid\; 0 \le \theta_1 \le 2\pi, \;\; 0 \le \phi_1 \le \pi\} \tag{3}$$

The second is the domain of measurement, which describes all possible orientations of the measurement axis $\mathbf{B}$,

$$\mathbf{S^2}_{mes} = \{(\theta_2, \phi_2) \;\mid\; 0 \le \theta_2 \le 2\pi, \;\; 0 \le \phi_2 \le \pi\} \tag{4}$$

[2] The operators R_{in} and R_{out} are a very simple way of mathematically visualizing the coordinate transformation of $\theta_{1,2}$ and $\phi_{1,2}$ into the Q-frame.

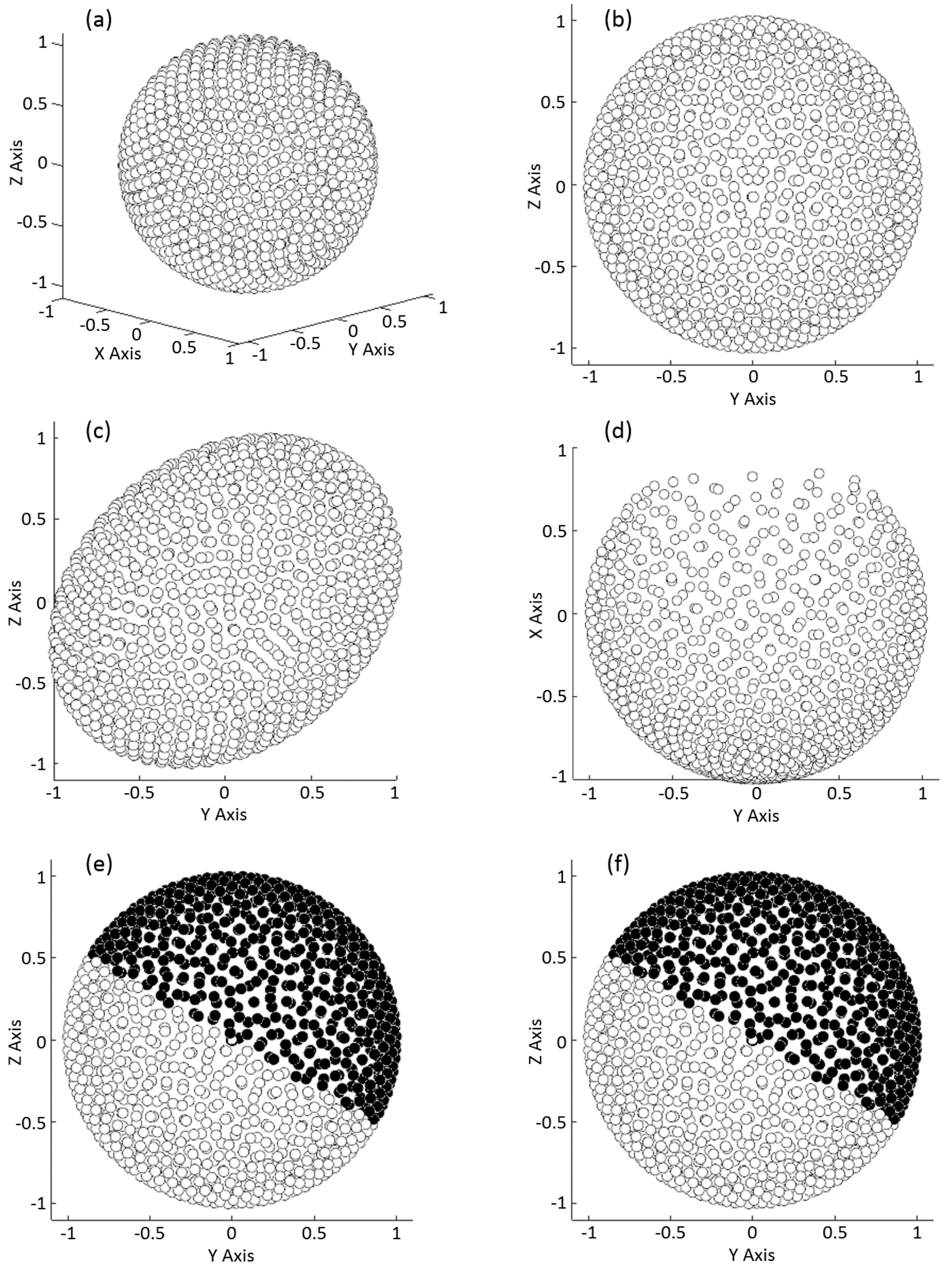


Figure 2. Simulated SNP measurement field. (a) and (b) Measurement field with no misalignment. (c) Misalignment applied by setting $\alpha_1 = 30^o$ (d) Misalignment applied by setting $\alpha_2 = 30^o$ (e) Misalignment applied by setting $\beta_1 = 30^o$ (f) Misalignment applied by setting $\beta_2 = 30^o$. In Figures (e) and (f) the data is segregated into two hemispheres with black and white face circles. This is to show clearly the effect of misalignment, which is identical in these two cases.

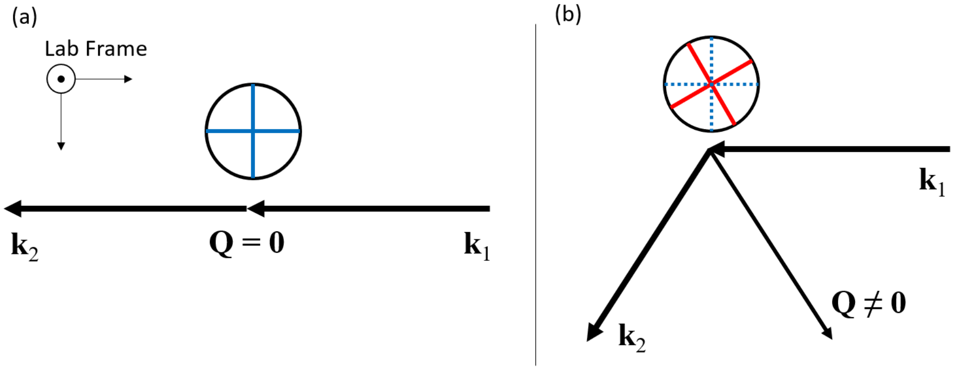


Figure 3. Orientation of the Q-frame for different values of $\mathbf{Q}$ as seen in the lab frame. (a) The transmission configuration during calibration and, (b) Crystal measurement. The blue cross (solid & dashed) indicates the Q-frame orientation for $\mathbf{Q} = 0$ and the red cross indicates the Q-frame orientation for $\mathbf{Q} \neq 0$.

Table 1. (Left) The opposing angles required for measuring the j^{th} vector component over the control domain. (Right) The opposing angles required for measuring the j^{th} vector component over the measurement domain.

	θ_2	ϕ_2
X	$\pi/2$	0
Y	0	$\pi/2$
Z	0	0

	θ_1	ϕ_1
X	$\pi/2$	0
Y	0	$-\pi/2$
Z	0	0

For each element of either $\mathbf{S^2}_{con}$ or $\mathbf{S^2}_{mes}$, Equation 2 will produce a vector $\mathbf{P} = \sum_j P_j \ \hat{\mathbf{e}}_j$ from three cardinal orientations corresponding to $j = 1$, 2 or 3 of the two opposing angles. The angles of these three opposing orientations for both the control and measurement domains are shown in Table 1. As an example, a measurement of some vector $\mathbf{P}_G$, from the control domain $\mathbf{S^2}_{con}$, first requires the setting $(\theta_1, \phi_1)_G$ to access the G^{th} point in that domain. Then one must independently measure the X, Y and Z components by setting (θ_2, ϕ_2) according to the Table 1 values for the $\mathbf{S^2}_{con}$ domain. When properly aligned, the resulting range of vectors, $\mathbf{P}$, for each domain ideally form a spherical field of radius P_o centered at the origin. Figures 2(a) and (b) illustrate a three-dimensional plot of the ideally aligned perfectly spherical measurement field resulting from the $\mathbf{S^2}_{mes}$ domain. In this context, misalignment can now be characterized by the Q independent phases α_1, β_1, α_2 and β_2, which add to the local control coordinates, i.e.,

$$P_j = \mathbf{B}_{xz}(\theta_2 + \alpha_2) R_{x_2}(\phi_2 + \beta_2) \cdot R_{x_1}(\phi_1 + \beta_1) R_{y_1}(\theta_1 + \alpha_1) \mathbf{P}_o \tag{5}$$

From Figure 2 one can see how the addition of these phases distorts the sphere. Figure 2(c) shows the resulting distortion in the measurement field (or equivalently the control field) when $\alpha_1 = 30^o$. From Figure 2(d) the distortion is entirely different when the same misalignment angle, $\alpha_2 = 30^o$, is applied. In contrast to these, Figures 2(e) and (f) show a tilting of the spherical field when the same 30^o misalignment is applied to either β_1 or β_2 separately. In all of these cases, switching the sign of the misalignment mirrors the distortion and, combinations of misalignment superpose. These phases can be deduced by an iterative comparison of measured data with simulation. The phases are then corrected for by electromechanical means within the hardware until the apparatus is well described by Equation 2 alone. We therefore define misalignment to be any distortion in either the control or measurement fields that is well described by Equation 2.

Physically, alignment through measuring either the control or measurement fields acts to maximize the transmission of polarization through the SNP apparatus along the cardinal directions of the Q-frame. The control and measurement fields, while both spherical, differ in that they describe different aspects of the SNP functionality. For example, consider Figure 3. Here, two values of $\mathbf{Q}$ in the lab frame are depicted, one for $\mathbf{Q} = 0$ and one for $\mathbf{Q} \neq 0$. For both values of $\mathbf{Q}$ the orientation of the corresponding spherical field is also depicted. The $\mathbf{Q} = 0$ case (Figure 3(a)) was used for alignment and is described by Equation 2. During the measurement of a crystal, $\mathbf{Q}$ will most likely be non-zero and the resulting Q-frame will be rotated relative to the Q-frame used for alignment (refer again to Figure 3(b)). Consider that at most six points of the measurement field are needed for any given $\mathbf{Q}$. In the lab frame, this amounts to the vertical poles and a great circle within the scattering plane of the measurement field. Practically, only these regions of the measurement field need to be well described by Equation 2. This is not so for the control field. Nuclear-magnetic scattering, at values of $\mathbf{Q} \neq 0$, would result in a scattered polarization directed away from the Q-frame cardinal axes [8]. If the control field is distorted in that direction and is fixed to the lab frame, then that distortion will propagate to any polarization measurement when the direction of scattered polarization coincides with that distortion, regardless of how spherical the measurement field is.

To emphasize the diference between the measurement and control fields we must first consider two other forms of distortion in addition to misalignment, noise and aberration. Noise we characterize as the expected variation in the spherical field due to the compounded Poisson error of the neutron detection process [9]. Aberration, in contrast, we define as distortion which is greater than the noise and is not well described by misalignment (i.e., Equation 2). As an example of noise, first consider Figures 4(a) and (c) which shows a simulation of 1024 discrete measurements on the $\mathbf{S^2}$ domain where Figures 4(b) and (d) are respective Mollweide projections of Figures (a) and (b). We use the Mollweide projection to visually retain the relative proportionality of the point distribution when compared to the three-dimensional plots [10]. In Figures 4(a) and (b) the average noise level is twice as large as the angular sampling period (i.e., the average angular distance between points in $\mathbf{S^2}$). Consequently, the noise is the dominant signal and the sampling period satisfies the Nyquist-Shannon sampling limit [11], meaning that this spherical field is noise limited. The noise level, displayed in Figures 4(c) and (d), has been reduced by an order of magnitude but the sampling period is unchanged (i.e., an equivalent 1024 points have been used). In this case, the noise fluctuation is small and features resulting from any other distortion may only be resolved if the Nyquist-Shannon sampling limit is met. This means that the spherical field for this reduced noise level is now Nyquist-Shannon limited. In other words, the angular resolution provided in Figures 4(c) and (d) is now adequate to resolve aberration in the spherical field occurring at large enough angular scales.

Now consider an example of aberration in the control field, for $\mathbf{Q} = 0$, depicted in Figures 5(a) and (b). We imagine this type of aberration being brought about by artifacts in either the laboratory or the sample environment. By mapping such an aberration, as in Figure 5(b), real space correlation to known artifacts would potentially allow the source of the aberration to be identified and removed. Figure 5 is a simulation where the noise level here is equivalent to that used in Figures 4(c) and (d). This simulated aberration appears in a decrease in polarization within the upper octant of the control field. Figure 5(b) shows that the aberration is localized within that octant and does not intersect the cardinal planes. If this aberration is now fixed relative to the lab frame then, for some value of $\mathbf{Q} \neq 0$, the cardinal planes would intersect it resulting in a loss of transmitted polarization. Figure 5(c) through (e) show the projections

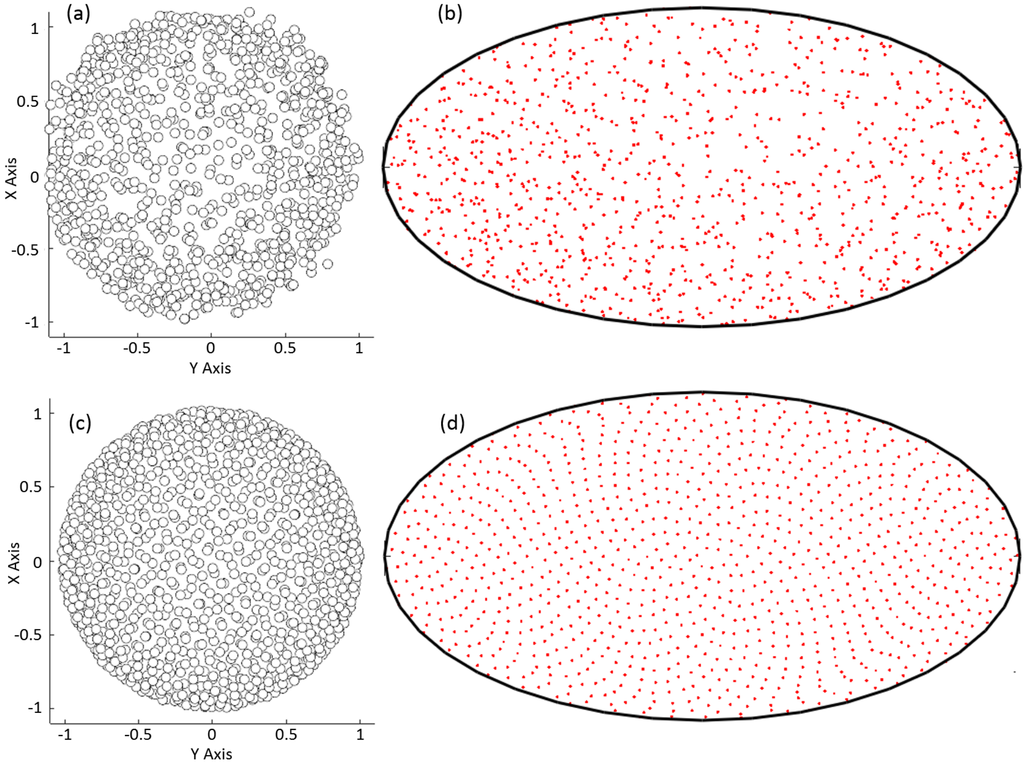


Figure 4. Simulation of the spherical field for two noise levels plotted in 3D and as a 2D Mollweide projection. (a) and (b) Noise Limited Data where the noise level is twice a large as the sampling period. (c) and (d) Nyquist Limited Data where the noise level is much smaller than the sampling period.

of this aberrated control field onto the cardinal planes for $\mathbf{Q} = 0$. Since the polarization in each projection follows the great circle (outlined in black) of that respective projection plane, the resulting measurement field will appear to be undistorted. Again, this is because only the undistorted X, Y and Z directed polarizations of the control field are used in constructing the measurement field. Therefore only a spherical measurement of the control field through the aberrated region will fully account for that distortion.

To characterize the type of aberration depicted in Figure 5 we now explore a strategy that involves expanding the measurement field in terms of a finite series of spherical harmonics. Techniques similar to this have been used to model three-dimensional data for Computer Generated Imaging (CGI) [12].

$$P(\theta, \phi) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \epsilon_{nm} Y_{nm}(\theta, \phi) \tag{6}$$

Equation 6 describes an expansion of the magnitude of the measured polarization on the $\mathbf{S^2}$ domain in terms of the spherical harmonics $Y_{nm}(\theta, \phi)$ with coefficients ϵ_{nm} for degree n and order m. Here the maximum degree, N, is determined by the total number of measurements

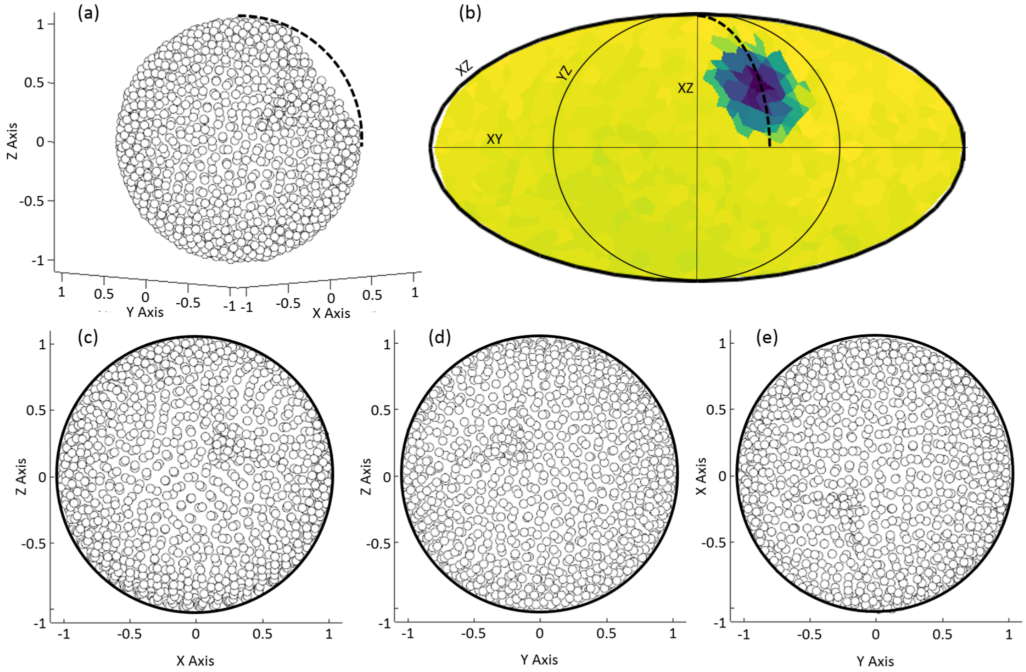


Figure 5. Simulated aberration. The noise level is equivalent to Figure 4(b). (a) A three-dimensional plot showing a decrease of polarization relative to the control field. The dotted line indicates the undistorted polarization. (b) Mollweide projection of of the simulation showing that the distortion is localized within one octant of the sphere. Here arcs and lines indicate great circles and are labeled according to the respective plane that contains them. The dotted line of (a) is also shown. (c-e) Projections of the simulation onto the planes labeled in (b). Great circles act as a guide for the eye.

taken, similar to discrete Fourier analysis.

To apply this expansion, we first begin by correcting for angular aberration. Angular correction serves to regularize the spherical field for harmonic decomposition by Equation 6. The operation we suggest here is a simplified form of image warping and we refer to Wolberg (1990) , for a more in depth treatment [13]. This is a procedure which first involves calculating the set of spherical angles, $\{A_k, B_k\}$, from the set of vectors, $\{\mathbf{P}_k\}$. Each A_k and B_k then correspond to the expected angles, a_k and b_k, set by the instrumentation such that the resulting angular aberration is the respective difference,

$$\delta_k = A_k - a_k, \qquad \rho_k = B_k - b_k \tag{7}$$

To later apply this to an arbitrary measurement we may linearly interpolate between angles such that for an arbitrary pair of measured angles, (γ, ζ), which, for example, are in the intervals, $A_k \leq \gamma \leq A_{k+1}$ and $B_k \leq \zeta \leq B_{k+1}$, the calculated aberration for those angles is,

$$\Delta_\delta = \Big|\frac{\gamma - A_k}{A_{k+1} - A_k}\Big|(\delta_{k+1} - \delta_k) + \delta_k \tag{8}$$

$$\Delta_\rho = \left|\frac{\zeta - B_k}{B_{k+1} - B_k}\right|(\rho_{k+1} - \rho_k) + \rho_k \tag{9}$$

The corrected angles for an arbitrary measurement are then,

$$\Theta = \gamma - \Delta_\delta \tag{10}$$

$$\Phi = \zeta - \Delta_\rho \tag{11}$$

With the angular distortion removed, correcting for radial distortion is more straight forward. This is accomplished by describing the set of magnitudes, $\{P_k\}$, of the calibration data with Equation 6. A correction to an arbitrary measurement, $P_{measured}$, with spherical angles Θ and Φ is then,

$$C(\Theta, \Phi) = \sum_{n=1}^{N} \sum_{m=-n}^{n} \epsilon_{nm} Y_{nm}(\Theta, \Phi) \tag{12}$$

such that,

$$P_{corrected} = P_{measured} - C \tag{13}$$

The corrected polarization described in Equation 13 is subject to the innate Poisson error of the calibration data, $\{\mathbf{P}_k\}$, and this error will propagate to the corrected polarization. We suggest that the error in $\{\mathbf{P}_k\}$ be kept to a level much smaller than that used for routine measurement.

3. Discussion

Thus far we have considered the limited case where both the control and measurement fields, measured at one value of **Q**, are fixed to the lab frame and are therefore Q-dependent. This dependency is described simply by a rotation of these fields about the Z-axis (Figure 3). In this sense, calibration requires characterization of one control and measurement field. Experimentally, we estimate that the measurement of a single spherical field would take about 20 hours for a neutron flux on the order of $2 \times 10^6/\text{cm}^2/\text{s}$. The simplicity of this kind of Q-dependency may be somewhat of an ideal since the modern SNP apparatus requires movement of instrumentation in the lab frame to measure different **Q**. An actual Q-dependency may be more complicated requiring field characterization at more than one **Q**. In any case, the strategy we have outlined for characterization of the control and measurement fields would not change.

Finaly, it must be pointed out that SNP calibration typically involves only measuring vector components along great circles within the control and measurement fields [4, 14, 5]. To date, there has been no attempt to characterize variation in an SNP apparatus over a spherical domain. As a result, it is not clear in what context aberration will significantly affect SNP measurements. For minor aberration in the control field, that is an aberration on the order of 2% above the noise, we expect aberration to have the potential of severely effecting measurements of weak nuclear-magnetic reflections. For these reflections, the nuclear component may be significantly stronger than the magnetic and, any loss in intensity could wrongfully suggest that the signal is purely nuclear.

Acknowledgments

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and

Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. The work utilized facilities supported in part by the National Science Foundation under Agreement No. DMR-1508249.

This work was supported by the Department of Energy, Basic Energy Sciences Project, Award DE-SC0016434.

References

[1] Brown P J, Nunez V, Tasset F, Forsyth J B and Radhakrishna P J 1990 *Phys.: Condens. Matter* **2** 9409
[2] Tasset F 1989 *Physica B* **156** 627
[3] Lelievre-Berna E et al. 2005 *Physica B* **356** 141-5
[4] Hutanu V, Meven M, Masalovich S, Heger G and Roth G 2011 *Journal of Physics: Conference Series* **294** 012012
[5] Takeda M, Nakamura M, Kakurai K, Lelievre-Berna E, Tasset F and Regnault L-P 2005 *Physica B* **356** 136
[6] Janoschek M, Klimko S, Gahler R, Roessli B and Boni P 2007 *Physica B* **397** 125
[7] Blume M 1963 *Physical Review* **133** 1366
[8] Brown P J 2006 *Neutron Scattering from Magnetic Materials* ed T Chatterji (The Netherlands: Elsevier) chapter 5 pp. 215-244
[9] Lelievre-Berna E, Brown P J, Tasset F, Kakurai K, Takeda M and Regnault L-P 2007 *Physica B* **397** 120-4
[10] Snyder J P *Map Projections - A Working Manual* (Washington: United States Government Printing Office) chapter 31 pp. 249-252
[11] Shannon C E 1998 *Proceedings of the IEEE* **86** 447-57
[12] Basri R and Jacobs D W 2003 *IEEE Transactions on Patter Analysis and Machine Intelligence* **25** 218-33
[13] Wolberg G 1990 *Digital Image Warping* (Wiley-IEEE Computer Society Press)
[14] Wang T, Parnell S R, Hamilton W A, Li F, Washington A L, Baxter D V and Pynn R 2016 *Review of Scientific Instrument* **87** 033901

META-DATA FOR IN-SITU MONITORING OF LASER POWDER BED FUSION PROCESSES

Shaw C. Feng[1], Yan Lu, and Albert T. Jones
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

ABSTRACT

Increasingly, a wide range of in-situ sensors are being instrumented on additive manufacturing (AM) machines. These sensors collect a variety of data that is used to monitor process performance and part quality. The amount, type, and speed of the collected data are unprecedented. Consequently, several data-related, standards issues are impeding the use of both data analytics and tool integration. Those issues include registration, curation, organization, storage, and management. This paper focuses on registration. It proposes the use of meta-data as a foundation for new interface and exchange standards. Standards that will facilitate the use of several types of in-situ sensors that monitor laser powder bed fusion (L-PBF) processes. The paper also includes an example data model that captures the properties and relationships among those meta-data. The data elements in that model provide industrial users with the capabilities and formats to capture, exchange, and share L-PBF process data.

1. INTRODUCTION

The value of parts fabricated using additive manufacturing (AM) continues to grow. Current market projections estimate that by 2025 that value will be approximately $7.65B in North America and $21.5B worldwide [7]. However, there are still hurdles impeding the widespread acceptance of AM as a reliable and cost-effective production technology. These hurdles are due primarily to the high variability in the quality of AM-produced parts. Quality problems include undesired pores or cracks, dimensional inaccuracies, poor surface finish, and inhomogeneous mechanical properties.

Factors causing quality problems include variable material properties, improper process settings, and the dynamic build environment. Many ongoing research efforts are being conducted to understand the impacts of those factors both individually and collectively. These efforts typically focus on two tasks: 1) identifying the material-process-structure-property relationships and 2) using those relationships to improve both process control and part quality. Both tasks rely on a variety of in-situ monitoring sensors for their success. That variety includes imaging sensors, thermal sensors, video cameras, acoustic sensors, ultrasonic sensors, and vibration sensors. The data coming from those sensors are characterized by volume, variety, velocity, and veracity [13]. A systematic organization of that data is needed for both real-time monitoring and off-line design and planning. The benefits of registering data are 1) accessing validated data with known time, locations, and approvals, 2) data alignment and fusion, 3) detecting defects traceable to process, material, equipment parameters, and 4) validating AM process models using validated data.

This paper describes our efforts implementing one important aspect of meta-data for in-situ monitoring of laser powder bed fusion processes. Meta-data are data types that describe how data were collected, including scanning strategy, sensor types, sensor configuration, calibration, and setup parameters. Meta-data is for AM data registration. AM data registration involves 1) a data curation process by which the context of the data (i.e., meta-data) is captured as well as 2) a unique data object identifier. This identifier can be used for generation of a persistent object identifier from a trusted registration authority. Registration is necessary to compare and fuse multi-sensor, in-situ monitoring data.

Most of this paper focuses on data collected by four types of sensors during a powder bed fusion (PBF) process. The four sensor types are imaging, thermometry, acoustic emission, and acceleration. Although profilometry is another possible sensor type, it still being studied. It will be added to our categorization when more detailed studies are available. Details in meta-data are in data elements in this paper.

Our proposed data elements approach is to link the in-situ data with product, process, and resource information. This paper also provides the XML representations of several data elements related to the raw monitoring data we collected by the sensors.

[1] Contact Author: shaw.feng@nist.gov

Feng, Shaw C.; Lu, Yan; Jones, Albert T. "Data Registration for In-Situ Monitoring of Laser Powder Bed Fusion Processes." Paper presented at International Mechanical Engineering Congress and Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

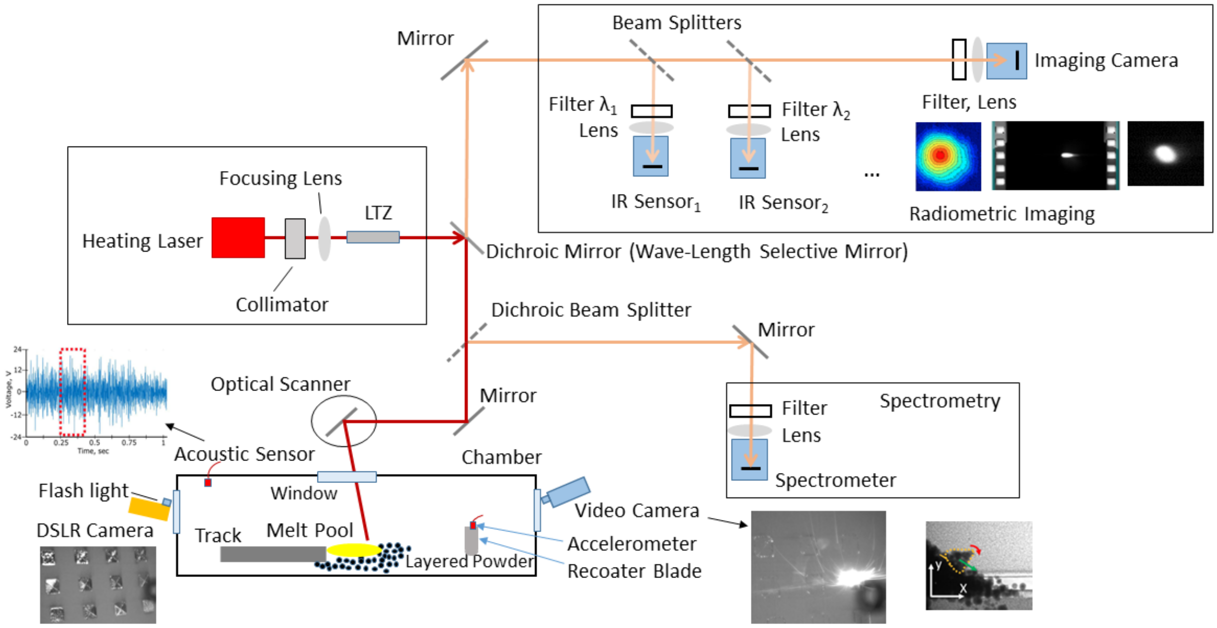


Figure 1 Schematic diagram of sensors for in-situ monitoring of L-PBF

These representations are key to 1) implementing and replicating our approach for data registration and 2) enabling the future standards on common data formats and data exchange.

The paper has six sections. Section 2 reviews related publications in in-situ monitoring. Section 3 describes data elements for in-situ process monitoring. Section 4 prescribes a data model based on the identified data elements to enable implementations. Section 5 provides an example and discuss the usage of the data model. Section 6 concludes the paper.

2. REVIEW OF IN-SITU MEASUREMENT RESEARCH

To develop a data model for data registration, we need to understand the current use of sensors for in-situ monitoring. The section provides a review of sensors, data fusion, sensor categorization, and research needs. Recently, researchers have been integrating L-PBF sensors and developing techniques to evaluate the data they provide.

2.1 Sensors for In-situ Process Monitoring

In-situ process monitoring, which is necessary for real-time control, is enabled by sensors and data analytics. In this paper, we focus on only four types of in-situ monitoring sensors: co-axial imagers, off-axis imagers, acoustic sensors, and accelerometers (see Figure 1). Co-axial sensors, which share the same axis as the laser beam, are optical sensors that can generate images of temperatures or melt-pool geometries. For the temperature measurement, a multi-wavelength pyrometer can provide more accurate temperature data. For the melt-pool geometry measurement, 2-D images are used to measure melt-pool dimensions. Moreover, a spectrometer can be used to analyze energy peaks and spatters in the melting process [2].

Off-axis sensors include Digital Single Lens Reflex (DSLR) cameras, acoustic sensors, and accelerometers. A DSLR camera can take images of the powder bed each time it is triggered. A combination of flashlights from different angles and illuminations can detect anomalies on each scanned layer [11]. A high-speed camera can record the laser-scanning process including melting, solidifying, and tracking. Acoustic sensors, which generate sound signatures in the frequency domain, can detect anomalies in the scanning process [15].

These different sensor types have different sensor capabilities. Lane et al. [9] characterized sensor capabilities using three metrics: spatial resolution, temporal resolution, and sensitivity. The authors' goal was to evaluate the capabilities of various sensors to determine how well they can detect defects and irregularities in AM processes. Yadroitsev et al. [19] described sensors used in the selective laser-melting process. The authors listed sensors to measure powder-material properties, process parameters, and their relationships to the instabilities in the process that can lead to defects. The same authors also characterized various defects on tracks and part surfaces.

Smith et al. [16] showed acoustic parameters used in spatially resolved, acoustic spectroscopy to detect near-surface defects such as pores, cracks, and voids. Depond et al. [3] developed a low-coherent, laser-scanning interferometry sensor and showed sensor parameters to measure powder layer-surface roughness. Bertoli et al. [1] estimated cooling rates using high-framerate video and multi-physics simulations. They reported consistency between the solidification shown in picture frames and the simulations. Heigel and Lane [8] measured melt-pool temperatures and dimensional characteristics using an infrared camera outside the chamber with temperature calibration. Hooper [9] demonstrated in-line measurements of melt-pool temperature and cooling rates using a coaxial laser and imaging design.

Fisher et al. [5] also used a coaxial sensor for monitoring melt pools. In the paper, the authors identified metrics for using the data from that sensor – including the cross-sectional area, the temperature changes, and the plate temperature. In addition to sensors for characterizing melt-pool temperatures, there are sensors for measuring individual parameters, such as melt-pool sizes. Tan et al. [17] listed specific material and process parameters to model the melting process. They also proposed a temperature-measuring technique using a pyrometer.

There is the fifth type of sensors: chamber environment sensors, e.g., inert gas flow meter, inert gas pressure sensor, and CO_2 concentration sensor. They are located outside the view shown in Figure 1 and not used to measure the melt pool or scanned layers.

2.2 Sensor Fusion In-Process Monitoring Data

As a new attempt at monitoring AM fabrication processes, researchers are now using several different sensor technologies simultaneously and fusing the resulting data. Foster et al. [6] used staring-video cameras and coaxial cameras to collect data for monitoring melt-pool characteristics. Additionally, in that same paper, the authors surveyed existing, ultrasonic, in-process sensors that can be used to detect pores and delamination. In a survey focused on direct energy-deposition processes, Reutzel et al. [14] described measurements of melt-pool geometry and temperature. The authors made geometric measurements based on images taken by a single-color camera in the infrared (IR) range. Temperature measurements, however, were based on images taken from a dual-color camera. The authors aligned the images with built-in reference marks in addition to the part design.

Everton et al. [4] described specific sensors and sensing techniques for monitoring part buildup, layer-by-layer, in L-PBF processes. Purtonen et al. [12] focused on optical sensors and acoustic sensors. Optical sensors included photodiodes, spectrometers, Charged Coupled Device (CCD) and Complementary Metal Oxide Semiconductor (CMOS) imaging sensors, acoustic sensors, pyrometers, and infrared cameras. Acoustic sensors included microphone and signal analysis software in the frequency domain for detecting anomalies in melting. Sensor fusion has been increasingly a research topic for better understanding of AM processes.

2.3 Sensor Categorization

In this paper, we aggregate currently available sensors developed for in-situ, L-PBF monitoring as shown in Tables 1 (a), (b), and (c). The leftmost column lists what is being monitored. Items in the list includes the entire melt pool, a freshly coated layer, a newly scanned layer, workpiece, chamber, and blade. The top row in each table lists available sensor types. The list includes still image cameras, video cameras, infrared cameras, pyrometers, thermocouples, sonic sensors, ultrasonic sensors, strain gages, accelerometers, CO2 sensors, air-pressure gages, and air-flow meters. These sensors belong to the following five types: photogrammetry, thermometry, acoustic emission, mechanical sensing, and chamber environment sensing.

Table 1 (a) Sensor categories and defect detection – photogrammetry

		Photogrammetry (still image or video)	
		Unstructured light (CMOS, CCD cameras)	Structured light (CMOS, CCD cameras)
Meltpool		Meltpool shape irregularities, e.g., key hole and size too small. Track irregularities, e.g., under melting, over melting, metal balls, and discontinuity	N/A
Layer	**Freshly coated**	Waviness, voids	Same as left with more contrast
	Scanned	Track discontinuity, cracks, voids, and spatters. Coated powder layer irregularities, e.g., streaks, waviness, and metal obtrusion	
Workpiece		N/A	N/A
Chamber		Plume, spatter, spark	N/A
Blade		N/A	N/A

2.4 Research Needs for Data Elements for Registration

Clearly, a variety of sensors are being used for in-situ monitoring of different AM processes. Those sensors provide a plethora of data, including images, video clips, temperature data, and acoustic signals. While the data from individual sensors are important, correlations among those data can be extremely valuable. Sensor data must be registered in a data repository providing metadata before the data can be applied for analysis and correlated for decision making. The data correlations are necessary to determine the state of the powder-fusion process, the material microstructure, and the fabricated part. For example, without correctly aligning measurements in the spatial and temporal domains, conflicting predictions can be made on the part quality

Table 1 (b) Sensor categories and defect detection – thermometry

		Thermometry				
		Radiometry			**Non-radiometry**	
		Infrared imaging IR camera with filter (still or video)	**Pyrometry**		**Thermocouple**	**Thermometer**
			Single bandwidth (Pyrometer)	**Multi-bandwidth (multi-bandwidth pyrometer)**		
Meltpool		Melt pool temperature profile	Melt pool temperature (with correction of emissivity) to detect under/overheating problems	Same as left, but calibrated with near true temperature black body	N/A	N/A
Layer	**Freshly coated**	N/A	Powder bed temperature	Temperature monitoring	N/A	
	Scanned	N/A	Powder bed temperature	Same as left		
Workpiece		N/A	N/A	N/A	Temperature monitoring	N/A
Chamber		N/A	N/A	N/A	N/A	Temperature monitoring
Blade		N/A	N/A	N/A	N/A	N/A

Table 1 (c) Sensor categories and defect detection – acoustic emission, mechanical sensing, and chamber environment sensing

	Acoustic emission		**Mechanical sensing**		**Chamber environment sensing**		
	Sonic (microphone)	**Ultrasonic (Ultrasonic sensor)**	**Strain (Strain gage)**	**Acceleration (Accelometer)**	**CO_2 concentration gaging**	**Air pressure gaging**	**Air flow metering**
Meltpool	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Layer	Sparking anomalies	N/A	N/A	N/A	N/A	N/A	N/A
Workpiece	Cracking	Crack, void detection	Thermal and residual stresses	N/A	N/A	N/A	N/A
Chamber	N/A	N/A	N/A	N/A	CO_2 concentration	Air pressure	Air flow
Blade	N/A	N/A	N/A	Waviness on the powder layer, Metal Obtrusion due to lamination	N/A	N/A	N/A

Metadata for capturing the ‘context’ of that data, both in the product development lifecycle and in the material and equipment supply chains, is also needed for qualification. That context should include three types of information related to individual sensors and their configuration, to the design, build, post-process, and inspection activities, and to material, equipment and personnel. This information is necessary for downstream applications, such as data analytics, and will enable users to analyze the data correctly. Data correctness reduces wrong decisions to be reached during and after L-PBF.

AM dataset registration as a research topic continues to expand. Nevertheless, the current limitations of that research are still impeding the use of advanced data analytics, which can be used to accelerate the understanding and control of AM processes. Current limitations include both a missing data identifier for tracing raw data objects and a lack of meta-data for data fusion and analytics. First, raw data objects from in-situ sensors are often stored in isolation: either on local computers, laboratory servers, or mobile drives. Since these data objects lack standardized syntax or semantics, each data object may not have a persistent identifier assigned. This makes it very difficult to trace the origins of the data objects where they are located. This, in turn, impedes the fusion of data objects collected from different monitoring systems and stakeholders.

Second, estimating the correlations among various sensor-data types is very difficult since there is no information to link the data correctly in both time and space. In the following sections, we present a data registration approach to contextualize AM in-situ data sets. The approach has the potential for standardization that will facilitate the integration of AM lifecycle data integration. Lastly, there is no contextual information for the product’s lifecycle. This is one of major barriers for part qualification and verification to ensure AM product quality.

Multi-modal, nondestructive sensors collect a variety of data objects for monitoring AM processes. Comparing, fusing, and correlating these data objects to the original AM design, the current microstructure, and the final part quality is a major problem. The quantity and quality of different data types are also causing curation, organization, and administration problems that can limit the usage of those data objects. In our view, data registration is focused on contextualizing data in time and space (see Figure 2) for capturing the meta information necessary.

Figure 2 shows a picture of how in-situ data can be registered geometrically onto a voxel model. The additional data necessary to do this are shown on the right-hand side of Figure 2. For example, the laser-scanning strategies are critical since the scanner-motion control and laser-control commands provide fundamental references to contextualize in-situ data sets both spatially and temporally. The rest of the Section describes data elements that are needed for registering in-situ monitoring data, including scanning paths, melt-pool monitoring (MPM) images and videos, layer-wise images, and signals in acoustic emission.

Unfortunately, registering spatial-temporal data is not enough to guarantee better process control. Control requires a transformation of the raw sensor and calibration data into the same coordinate system. Control also requires linking all in-situ data to the design, geometry, tolerance, material, and inspection information.

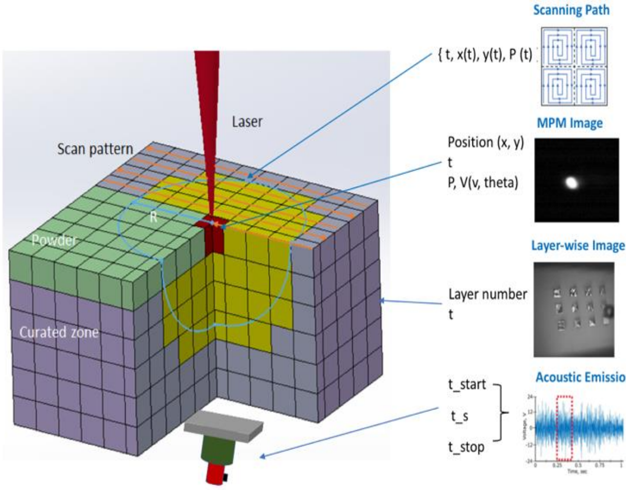


Figure 2 Registering AM data using a voxel model

3 Data Elements for In-situ Process Monitoring

3.1 Laser-scanning-strategy Data Elements

Laser-scanning strategies can provide both temporal and spatial references for the spatiotemporal alignment of in-situ measurements from different sensors. The scanning-strategy registration method is process-oriented, based on the xy2-100 protocol used in some open AM systems. The xy2-100 files provide the laser-spot positions, laser power, and camera-trigger timing [20]. This method is used primarily to register the position and the time when the image is taken by a camera. Required data elements are shown in Table 2. The scan starting time can be used as the temporal reference to align the data in the time domain, e.g., acoustic data.

Table 2 Process-oriented Data Elements

Data Element	Description
Build ID	unique identification of the associated build
Part ID	unique identification of the associated part
Scanning ID	unique identification of the data element
Layer number	the powder layer number
Command time (t)	the time that a position command is sent
Scan position (x, y)	the commanded location of the laser beam
Laser power (P)	the power of the laser beam
Scanning speed (V)	the scanning speed of the laser beam

The metadata shown in Table 2 can be provided by an open-architecture AM system, e.g., the AM Metrology Test Bed from the National Institute of Standards and Technology[2]. For commercial AM systems, part-oriented scanning strategies are used to scan the layer with minimized scanning time with acceptable part quality. The strategy specifies scan paths and process parameters based on partitions of the layer and is vendor-dependent and proprietary, so the registration of this kind of scan strategies is not discussed here.

3.2 Layer-wise images

Layer-wise images are dominating the current datatypes used in monitoring AM processes. There are two methods for registering layer-wise images: an individual image or a folder of images. The two methods are described below.

3.2.1 Registering an Individual File

When there is a limited number of individual images on a pre-scanned or post-scanned powder layer, each image can be registered with the data elements in Table 3 to provide the necessary meta information.

Table 3 Image Data Elements

Data Element	Description	
Image (or Movie) Name	the name of the image or movie	
Image (or Movie) ID	a unique identification (ID) number of the image or movie	
Build ID	the ID of the build for the part	
Layer number	the layer number of this image	
Time	the time it was taken	
Folder Path	the directory path for locating the folder that this image or movie was saved	
Flash condition	if flash lights are used, a description of the flash light angle relative to the layer	
Sensor ID	the identification of the image sensor	
Sensor description	sensor type (e.g., InSb, CMOS, Photoiode), purchase data, wavelength ranges, lens distortion information, and other specifications, including filters.	
Sensor configuration ID	Sensor configuration description must have the following tags	
	Original window size in pixels	mm x mm
	Cropped	(y/n)
	Pixel pitch	(µm/pixel)
	Magnification	magnification factor
	Viewing angle	(degree)
	Bit depth	the number of levels on grayscale or color scale
	Shutter Speed	the amount of time that the shutter is open for taking an image (s)
	Optical filter bandwidth	minimum and maximum wavelengths in nm
	Sensor calibration information	the date of calibration, the method of calibration, and person who performed the calibration

3.2.2 Registering a Folder of Images

When there is a large number of individual images on a pre-scanned or post-scanned powder layer, the images may be organized in a file folder. In this case, users should register all the folders with all the optical images from the first layer to the last layer. Data elements are in Table 4.

Table 4 Image Folder Data Elements

Data Element	Description
Build ID	the ID of the build
Folder path	the path of the file directory of the folder
Layer range	the start layer and the end layer included in the folder
Start time	the time that the first image was taken
Stop time	the time that the last image was taken
Flash configuration enumeration	the enumeration of the index of flash lights and their configuration descriptions
Image prefix	the prefix of an image in the folder
Flash condition	the condition of an indexed flash light
Sensor ID	the ID of the sensor
Sensor description	the description should include the type of sensor, the purchased date, sensor specification, and the information on lens distortion
Sensor installation	the view angle
Sensor setting and configuration	description of settings and configuration of the sensor
Configuration ID	ID of a sensor configuration
Image size in both X and Y directions	pixel by pixel
Cropped	(y/n)
Pixel pitch	(µm/pixel)
Magnification	the magnification factor
Viewing angle	the angle relative to the build plate normal
Bit depth	the number of levels in grayscale or color scale
Shutter Speed	the amount of time that the shutter is open for taking an image (s)
Optical Filter Bandwidth	minimum and maximum wavelength in nm
Sensor calibration information	the description of sensor calibration

3.3 MPM Data Elements

In this section, we focus on registering still images and video files since they are used by metrologists for process monitoring.

[2] https://www.nist.gov/el/ammt-temps

Since melt-pool monitoring often requires high-speed cameras, which can capture thousands of images per layer, we recommend users register images in a folder. Data elements are in Table 5.

Table 5 An Image in the Folder (Data Elements)

Data Element	Description	
Image Name	the name of the image	
Image ID	a unique identification number of the image	
Build ID	the ID of the build	
Triggering Time	the time when the camera is being triggered to take the picture based on the scanning program in xy2-100	
Instant laser power	the laser powder (W) at the time of image taken	
Frame Rate	frame per second (fps) if it is a movie	
Folder Path	the directory path for locating the folder that this image was saved	
Sensor ID	the identification of the image sensor	
Sensor description	sensor type, purchase data, specifications, lens distortion information, etc.	
Sensor installation	installed date, installer	
Sensor configuration ID	Sensor configuration description must have the following tags	
	Original window size in pixels	mm x mm
	Cropped	(y/n)
	Pixel pitch	(nm/pixel)
	Magnification	magnification factor
	Bit depth	the number of levels in grayscale or color scale
	Shutter Speed	the amount of time that the shutter is open for taking an image (s)
	Optical filter bandwidth	minimum and maximum wavelengths in nm
	Sensor calibration information	The date of calibration, the method of calibration, person who performed the calibration, and the calibration data

3.4 Acoustic Emission Data Elements

Acoustic sensors can capture the sparking and cracking sound generated in L-PBF. That sparking sound is generated by laser heating. Metal cracking is due to thermal stress. The acoustic emission data is a time series of signals. The file elements can be found in Table 6.

Table 6 Acoustic Emission Data Elements

Data Element	Description
File name	the name of the file
Uniquely generated ID	ID of the data
Start Time	Time of recording started
Stop Time	Time of recording stopped
Sampling interval	Time interval between two samples
Sensor location	The location of the acoustic sensor in the chamber
Setting	Acoustic sensor setting information

3.5 In-situ Measurement Uncertainty Quantification

There is always uncertainty in the data collected from sensors. Knowing the sources of that uncertainty is critical to applications such as data analysis and model validation. Uncertainty sources can be registered in categories in Table 7.

Table 7 Uncertainty Sources

Sensor Type	Uncertainty Source
Camera	• view angle variation due to installation • magnification factor variation due to the viewing angle • instantaneous Field Of View (iFOV) due to viewing angle • FOV due to viewing angle • Variation in the focus of lens
Laser spot	• The location relative to the build plate coordinates in the X and Y directions
Galvo Scanner	• The actual laser spot position in the X and Y directions relative to the build plate coordinates in the X and Y directions can deviate from the command position.
Image	• The laser (spot) is moving while the camera is taking a picture. Melt pool keeps changing during the exposure. Uncertainty is embedded in the shape and size of the measured melt pool.

4 Data Model for Meta-Data

To implement the identified data elements in Section 3, we developed a data model using the XML Schema language [18]. XML Schema was chosen based on the following reasons: (1) data structures in XML enabling efficient search/query, (2) predefined data types satisfying a variety of our modeling needs, (3) XML files validation using XML schema, and (4) available tools for implementation. The user community can implement the data registration method using XML tools with the schema described in this Section. Our choice does not imply that XML Schema is the best language for data modeling. Every modeling language has its capabilities and shortcomings.

Feng, Shaw C.; Lu, Yan; Jones, Albert T. "Data Registration for In-Situ Monitoring of Laser Powder Bed Fusion Processes." Paper presented at International Mechanical Engineering Congress and Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

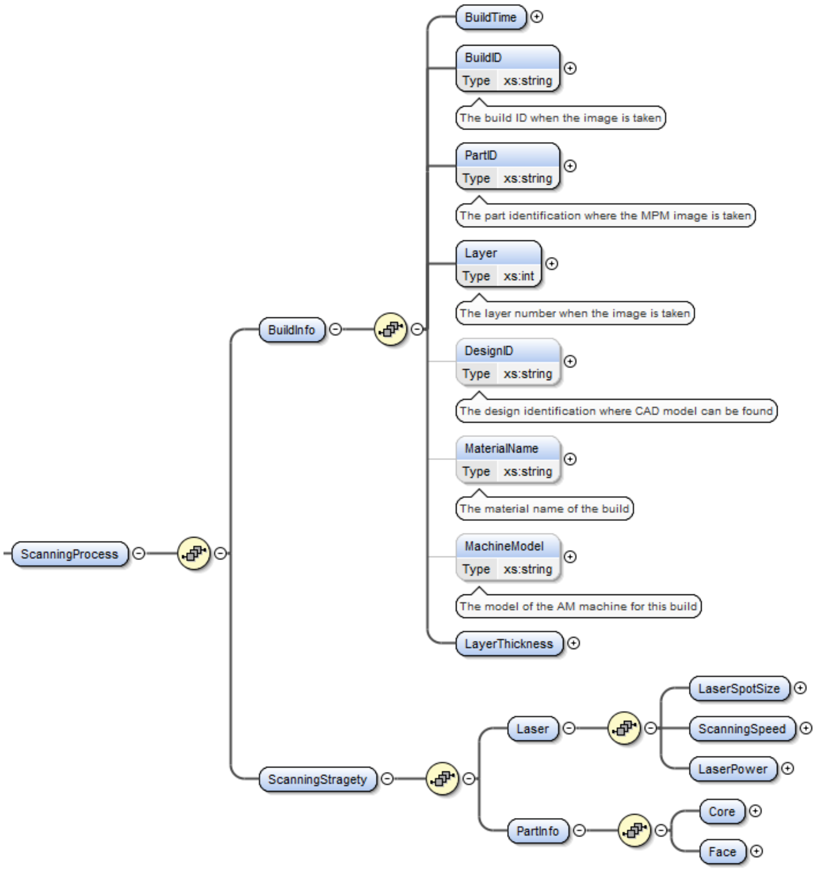

ScanningProcess
BuildInfo
BuildTime
BuildID
Type xs:string
The build ID when the image is taken
PartID
Type xs:string
The part identification where the MPM image is taken
Layer
Type xs:int
The layer number when the image is taken
DesignID
Type xs:string
The design identification where CAD model can be found
MaterialName
Type xs:string
The material name of the build
MachineModel
Type xs:string
The model of the AM machine for this build
LayerThickness
ScanningStragety
Laser
LaserSpotSize
ScanningSpeed
LaserPower
PartInfo
Core
Face

Figure 3 Data schema for scanning process

The data elements and attributes in the schema are based on the classification and data elements described in Section 4. In-Situ_Monitoring is the root element. It has four sub-elements: Scanning_Process, Layer_Imaging, Melt_Pool_Imaging, and Acoustic_Singnals. In the paper, we only describe a group of data elements for scanning process (Section 5.1) and another group for layer imaging (Section 5.2) as two examples. We create attributes to uniquely identify the data element. Note that attributes are not shown due to the limitation of the paper length. We create sub-elements to describe details of the data element.

4.1 Scanning Process Data Elements

As described in Section 3.1, a scanning process is for a specific build with a scanning strategy. The main data element is Scanning_Process, and it thus has two sub-elements: Build_Info and Scanning_Strategy. Figure 3 shows a graphical representation of Scanning_Process elements and its two sub-elements. Build_Info has two attributes, BuildID and Total_Number_of_Layers (not shown). Two sub-elements, Build_Time and Layer_Thickness provide detailed descriptions of Build_Info. Similarly, Scanning_Strategy has attributes to uniquely describe the data element. Three sub-elements to provide detailed information on Scanning_Strategy: Laser, Part_info, and G-code. Furthermore, Laser and the Part_info have attributes and sub-elements.

4.2 Layer Imaging Data Elements

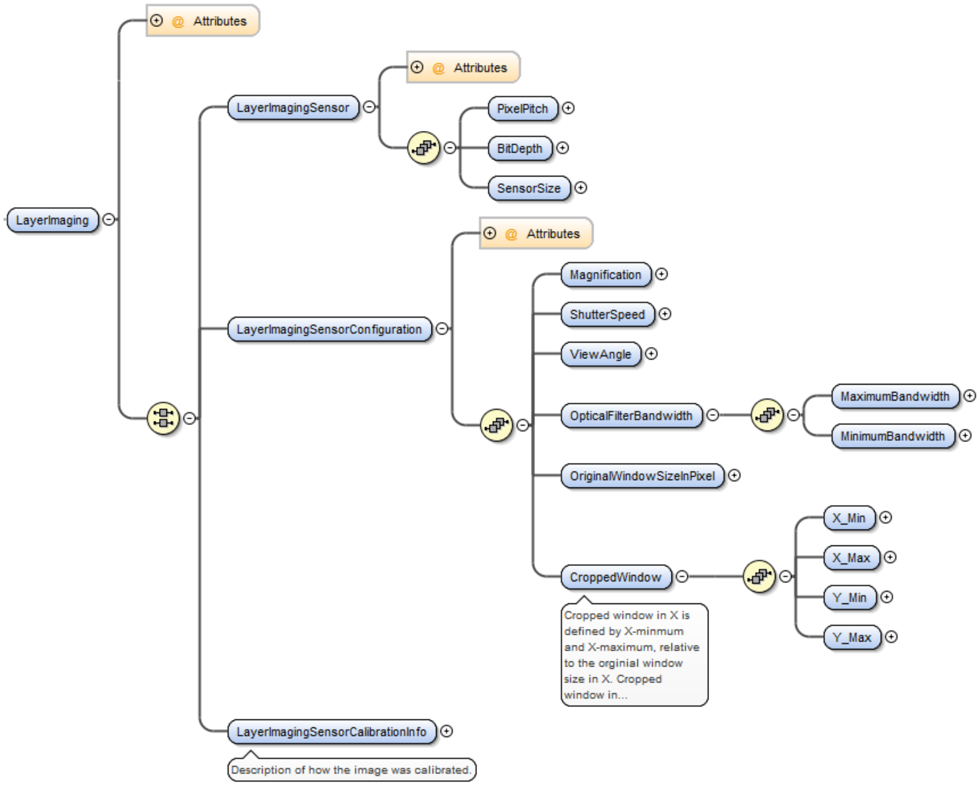
As described in Section 4.2, Layer_Imaging is the main element for layerwise images registration and six attributes and three sub-elements, as shown in Figure 4. Like the description in the previous section, attributes uniquely describe a data element. The data element can have one or more sub-elements that provide detailed information about the data element. Six attributes (not shown) are Image_Name, Image_ID, Time_Taken (the time when the image was taken), Folder_Path (the path to locate the folder where the image is saved), Flashing_Condition (a description of flash lighting, including number flash lights and their locations and flashing directions), and Still_Image_of_Movie (to indicate the registered item is a set of images or a movie).

Layer_Imaging_Sensor is one of three sub-elements and has two attributes and three sub-elements. Two attributes are Sensor_ID and Sensor_Description. Three sub-elements are Pixel_Pitch, Bit_Depth, and Sensor_Size. They all have quantity and unit as their attributes. Layer_Imaging_Sensor_Configuration has three attributes (not shown) and six sub-elements. Three attributes are Sensor_Configuration_ID, Sensor_Configuration_Description, and Cropped_or_not (for indicating whether the image is cropped or not). Six sub-elements are Magnification (the magnification factor of the lens), Shutter_Speed (camera shutter speed), View_Angle, Optical_Filter_Bandwidth, Original_Window_Size_in_Pixel, and Cropped_Window (an indication whether the window is cropped or not). Layer_Imaging_Sensor_Calibration_Info (to provide relevant information on the calibration of the sensor).

5 Sample Use Case

The data schema presented in Figure 3 and 4 were instantiated to capture the meta data which describe the scan command and in-situ tower camera based layerwise monitoring data set generated from a 3d build using an open architecture powder bed fusion system- NIST AMMT[3]. A Python program is created to register the layerwise image against the galvo position. Figure 5 shows a function diagram of the program.

[3] https://www.nist.gov/el/ammt-temps


Attributes
LayerImaging
LayerImagingSensor
Attributes
PixelPitch
BitDepth
SensorSize
LayerImagingSensorConfiguration
Attributes
Magnification
ShutterSpeed
ViewAngle
OpticalFilterBandwidth
MaximumBandwidth
MinimumBandwidth
OriginalWindowSizeInPixel
CroppedWindow
Cropped window in X is defined by X-minmum and X-maximum, relative to the orginial window size in X. Cropped window in...
X_Min
X_Max
Y_Min
Y_Max
LayerImagingSensorCalibrationInfo
Description of how the image was calibrated.

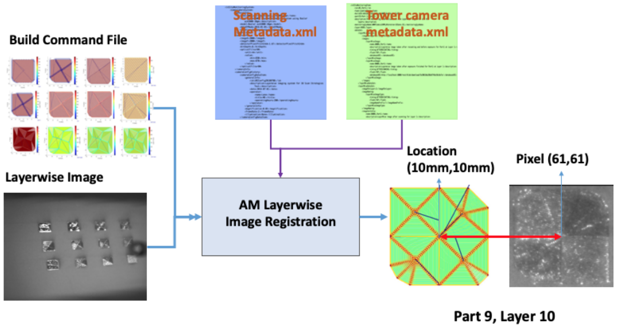


Figure 5 Use case for XML-based meta data

The tower camera metadata.xml file provides camera intrinsic and extrinsic parameters to remove the projection from the layerwise images and convert the image to the build platform coordinates. The Scanning Metadata.xml deciphers the build command data from which the contours of each part are extracted and used to segment the parts and convert the pixels of the image into galvo positions. The output of the functions are the registered after-exposure images of each part. Figure 5 shows the registered Part 9 layerwise image at Layer 10. Pixel (61, 61) corresponding galvo location at (10mm, 10mm).

6 Conclusion and Future Work

The use of laser powder bed fusion, L-PBF, technology to fabricate complex, metal parts in several industries has been steadily growing. As a result, the demands on both the quality and reliability of those parts have increased. Academic researchers and real-world manufacturers have implemented in-situ sensors to monitor L-PBF processes and to detect potential anomalies in the part.

The target audience of the paper are the practitioners who rely heavily on standards, interfaces and tools to integrate different systems and data. The work in this paper enables the future standards on common data exchange formats which will play critical role in implementing the scientific results from the academia and integrating and analyzing AM sensor data.

This paper focuses on defining a key element of those standards: meta-data. Specifically, this paper proposes new data elements, which can be used to characterize the properties and relationships among the data types captured by those sensors. Characterizing those properties and relationships requires a schema model; we have included a small number of examples of such a model.

Future work will be in three areas. One is to include other types of in-situ sensors emerging in the future, such as pyrometer and ultrasonic sensor. Second is to extend the number of schema model, standardize the resulting models, and integrate them with other existing schemas, such as powder material and machine schemas. We expect that these standards will lead to better implementations in the L-PBF user community. Third is to characterize data and meta data on image calibrations and define data elements. Specifically, a data model on methods and calibration instruments needs to be included in the developed data model. The data model will enable users of the data to compensate distortions in an image and quantify uncertainties in measured data.

DISCLAIMER

The work presented in this document is an official contribution of the National Institute of Standards and Technology (NIST) and not subject to copyright in the United States. Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by NIST. Nor does it imply that the products identified are necessarily the best available for the purpose.

REFERENCES

[1] Bertoli, U., Guss, G., Wu, S., Matthews, M., and Schoenung, J., "In-situ Characterization of Laser-powder Interaction and Cooling Rates Through High-Speed Imaging of Powder Bed Fusion Additive Manufacturing," Journal of Materials and Design, Vol. 135, 2017, pp. 385 – 396.

[2] Caelers, M., "Study of in-situ monitoring methods to create a robust SLM process," Thesis, KTH Royal Institute of Technology, 2017.

[3] DePond, P., Guss, G., Ly, S., Calta, N., and Deane, D., "In Situ Measurements of Layer Roughness During Laser Powder Bed Fusion Additive Manufacturing Using Low Coherence Scanning Interferometry," Journal of Materials and Design, Vol. 154, 2018, pp. 347 – 359.

[4] Everton, S., Hirsch, M., Stravroulakis, P., Leach, R., and Clare, A., "Review of In-situ Process Monitoring and In-situ Metrology for Metal Additive Manufacturing," Journal of Materials and Design, Vol. 95, 2016, pp. 431 – 445.

[5] Fisher, B., Lane, B., Yeung, H., Beuth, J., "Toward Determining Melt Pool Quality Metrics Via Coaxial Monitoring in Laser Powder Bed Fusion," Manufacturing Letters, Vol. 15, 2018, pp. 119 – 121.

[6] Foster, B., Reutzel, E., Nassar, A., Hall, B., and, Dickman, C., "Optical, Layerwise Monitoring of Powder Bed Fusion," Proceedings of the 26th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, 2015, pp. 295 – 307.

[7] Frost & Sullivan's Global 360° Research Team, "Additive Manufacturing Market, Forecast to 2025," MB74-10, Frost & Sullivan, Mountain View, CA, May 2016.

[8] Heigel, J. and Lane, B., "Measurement of The Melt Pool Length During Single Scan Tracks In a Commercial Laser Powder Bed Fusion Process," Proceedings of the ASME 2017 12th International Manufacturing Science and Engineering Conference, Los Angeles, CA, 2017, paper number: MSEC2017-2942.

[9] Hooper, P., "Melt Pool Temperature and Cooling Rates in Laser Powder Bed Fusion," Journal of Additive Manufacturing," Vol. 22, 2018, pp. 548 – 559.

[10] Lane, B., Grantham, S., Yeung, H., Zarobila, C., Fox, J., "Performance Characterization of Process Monitoring Sensors on the NIST Additive Manufacturing Metrology Testbed," Proceedings of the Solid Freeform Fabrication Symposium, 2017.

[11] Petrich, J., Gobert, C., Phoha, S., Nassar, A, and Reutzel, E., "Machine Learning for Defect Detection for PBFAM Using High Resolution Layerwise Imaging Coupled With Post-Build CT Scans," Proceedings of the 28th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, 2017, pp. 1363 – 1381.

[12] Purtonen, T., Kalliosaari, A., and Salminen, A., "Monitoring and Adaptive Control of Laser Processes," Physics Procedia, Vol. 56, 2014, pp. 1218 – 1231.

[13] Razvi, S., Feng, S., Lee, Y.-T., Witherell, P., "A Review of Machine Learning Applications In Additive Manufacturing," Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, paper number IDETC/CIE2019, August 2019, Anaheim, CA.

[14] Reutzel, E. and Nassar, A., "A survey of sensing and control systems for machine and process monitoring of directed-energy, metal-based additive manufacturing," Rapid Prototyping Journal, Vol. 21 Issue: 2, 2015, pp.159-167, https://doi.org/10.1108/RPJ-12-2014-0177.

[15] Shevchik, S., Kenel, C., Leinenbach, C., and Wasmer, K., "Acoustic Emission for In Situ Quality Monitoring in Additive Manufacturing Using Spectral Convolutional Neural Networks," Journal of Additive Manufacturing, Vol. 21, 2018, pp. 598 – 604.

[16] Smith, R., Hirsch, M., Patel, R., Li, W., Clare, A., and Sharples, S., "Spatially Resolved Acoustic Spectroscopy for Selective Laser Melting," Journal of Materials Processing Technology, Vol. 236, 2016, pp. 93 – 102.

[17] Tan, H., Chen, J., Zhang, F., Lin, X., and Huang, W., "Estimation of Laser Solid Forming Process Based on Temperature Measurement," Journal of Optics and Laser Technology, Vol. 42, 2010, pp. 47 – 54.

[18] XML Schema Definition Language, World Wide Web Consortium (W3C), 2012, https://www.w3.org.

[19] Yadroitsve, I., Bertrand P., and Smurov, I., "Parametric Analysis of The Selective Laser Melting Process," Journal of Applied Surface Science, Vol. 253, 2007, pp. 8064 – 8069.

[20] Yeung, H., Neira, J., Lane, B., Fox, J., and Lopez, F., "Laser path planning and power control strategies for powder bed fusion systems," In Proceedings of the 27th Annual International Solid Freeform Fabrication Symposium, Austin, TX, 2016, pp. 113–127.

Feng, Shaw C.; Lu, Yan; Jones, Albert T. "Data Registration for In-Situ Monitoring of Laser Powder Bed Fusion Processes." Paper presented at International Mechanical Engineering Congress and Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

**Proceedings of the ASME 2019
International Mechanical Engineering Congress and Exposition
IMECE2019
November 11-14, 2019, Salt Lake City, UT, USA**

IMECE2019-11345

PHYSICS–BASED SIMULATION OF AGILE ROBOTIC SYSTEMS

Pavel Piliptchak*, Murat Aksu, Frederick M. Proctor, John L. Michaloski
National Institute of Standards and Technology
Gaithersburg, MD

ABSTRACT

Development and testing of industrial robot environments is hampered by the limited availability of hardware resources. Simulations provide a more accessible and readily modifiable alternative to physical testing, but also require careful design to maximize their fidelity to the real system. In this paper, we describe new progress in building an entirely physics-driven simulation of the Agility Performance of Robotic Systems (APRS) Laboratory at the National Institute of Standards and Technology (NIST). To maximize the accuracy of physics-based interactions in this environment, we develop several general-purpose improvements to the simulation of parallel grippers and multi-object collisions. We use pick-and-place tasks from the APRS Laboratory as well as generic benchmarks to verify the performance of our simulation. We demonstrate that our proposed improvements result in more physically-consistent simulations compared to standard implementations, regardless of the choice of physics engine or simulation parameters.

Keywords: robotics, agility, grasping, pick-and-place, simulation

1 INTRODUCTION

Simulations are invaluable tools in developing industrial robotic systems. In situations where hardware resources are limited or costly to use, simulations can provide a test-bench for developing controllers, debugging software, and testing various work-space configurations [1]. One of the limiting factors to using a simulation in place of a real system is how well it replicates the physical dynamics of the real system. This discrepancy has inspired the development of new rigid-body physics engines that tackle specific problem areas, such as numerically stable contacts and joint articulations [2–4].

However, existing robotics simulators are not a one-size-fits-all solution. A significant amount of effort needs to be put into defining contact behavior, inertia matrices, friction coefficients, and numerical solver parameters to ensure acceptable performance [5]. This effort becomes increasingly laborious as the simulated environment becomes more complex.

In this work, we use the Robot Operating System (ROS) [6] and the Gazebo [7] robotics simulator to build an entirely physics-based simulation of the APRS Laboratory, an environment designed for evaluating the capabilities of robots performing industrial pick-and-place tasks [8]. This environment is fairly difficult to replicate with standard simulators due to the large number of small, low-mass objects that the robot must accurately manipulate. Along the course of building this simulation, we developed several practical and general-purpose improvements to simulating two important components of our pick-and-place task (namely parallel grippers and multi-object collisions).

This work presents the following contributions:

1. A complete simulation of the APRS Laboratory that integrates with existing APRS software
2. A simple controller for simulated parallel grippers that improves grasp stability
3. A method for simplifying contacts between many relatively stationary objects in simulation

[*] Contact author: pavel.piliptchak@nist.gov

Piliptchak, Pavlo; Aksu, Murat; Proctor, Frederick; Michaloski, John L. "Physics-based Simulation of Agile Robotic Systems." Paper presented at International Mechanical Engineering Congress & Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

In the following sections, we will (1) discuss relevant experiments in physics-based simulation at NIST and abroad; (2) describe the design of our APRS Laboratory simulation; (3) describe our two improvements mentioned above; and (4) present experiments verifying the benefits of these improvements.

2 RELATED WORK

Despite the popularity of Gazebo for robot simulation, there have been comparatively few publications that use it for physics-based grasping and manipulation—particularly in full-fledged, realistic environments. The most notable example is the Defense Advanced Research Projects Agency (DARPA) Virtual Robotics Challenge (VRC) which required grasping as part of several subtasks [9, 10]. Simulations have also been used for the dexterous manipulation of cloth [11], pick-and-place tasks [12], and dexterous reinforcement learning agents [13]. The robots in these simulations generally use high degree-of-freedom (DOF), anthropomorphic hands. They also generally handle palm-fitting objects rather than smaller objects that require fingertip-only grasps. In contrast, the APRS simulation robots use parallel grippers and manipulate peg-like objects with diameters that are under 1 cm.

Another angle of research has been verifying the physical accuracy of the physics engines themselves [14–16] with some works emphasizing grasping in particular [17, 18]. Often, this research seeks out general benchmarks that measure either: 1) how the engine's compute speed scales with simulation complexity; 2) how well the engine conserves energy/momentum; or 3) how far the engine deviates from some ground-truth behavior for a simple dynamical system. We verify the accuracy of the APRS simulation using similar methods (to be discussed in the Experiments section), but focus on benchmarks that are relevant to grasping and pick-and-place tasks. We do not make claims about the general accuracy of the tested physics engines.

NIST has also pursued research in verifying the physical accuracy of robot simulators [19, 20]. This earlier work focused on the simulation of the Talon Robot using the older Karma Physics Engine. The evaluation metrics used in this work were based primarily on mobility with limited discussion of grasping. In contrast, our work simulates a pick-and-place task using Gazebo's variant of the Open Dynamics Engine (ODE) [2, 5]. It focuses more finely on grasp stability and simulator performance.

3 SYSTEM OVERVIEW

In this section, we give a high-level overview of our APRS simulation. We will begin by reviewing the objectives and design of the physical APRS Laboratory. Following this, we will present the design of our simulation along with its interfaces to the existing APRS architecture.

APRS Architecture

The APRS Laboratory seeks to study and measure the "agility" of robotics systems, which widely refers to a system's retasking capability, robustness to failures, and interoperability [8]. As part of this research effort, the APRS Laboratory implements an archetypal agile robotics system that operates on a simple pick-and-place kitting task.

The system consists of a Fanuc LR-Mate 200iD and a Motoman SIA20F robot. These two robots are positioned on opposite sides of a conveyor, which carries several part trays containing gears. These gears must be moved by the robots (both equipped with parallel gripper end-effectors) to a specified kit tray. As the robots perform the kitting task, the system experiences several mock failures, such as a robot breaking down or a part being misplaced.

Reacting to these failures requires an agile approach to sensing, planning, and control—the details of which can be found in [8]. To briefly summarize:

1. The APRS Laboratory maintains an ontology describing relevant kitting objects and concepts
2. An overhead camera system is used to locate instances of objects described in this ontology
3. A planner uses the ontology and instance information to generate a high-level action plan
4. Low-level controllers execute this plan on the robots

In the described architecture, replacing the physical system with a simulation requires three steps: replicating the camera system, replicating the robot grasping dynamics, and creating an interface between the high-level plan and the low-level virtual controllers used by the simulated robots. Virtual sensors were explored in [21], while implementing a shared interface and achieving accurate dynamics is this work's focus.

Simulation Framework

The APRS simulation framework is built primarily using ROS, which is an open-source suite of robotics-related software packages that are interconnected through a shared message-passing framework. Out of the box, ROS packages provide us with the kinematic descriptions of the robots, joint controllers, joint-state information, forward/inverse kinematic solvers, and collision-aware trajectory planning. ROS can also be extended by writing new packages, which is how we implement the kinematic descriptions of the conveyor and parallel grippers in the system.

We use Gazebo, an open-source robotics simulator, to dynamically simulate the system. Gazebo is designed to be compatible with many of ROS's components, including the robot kinematic descriptions, joint controllers, and joint-state publishers. Gazebo also supports multiple physics engines and allows users to tune the behavior of each engine's numerical solver. We will provide a detailed discussion our choice of physics parameters and physics engine in the "Experiments" section.

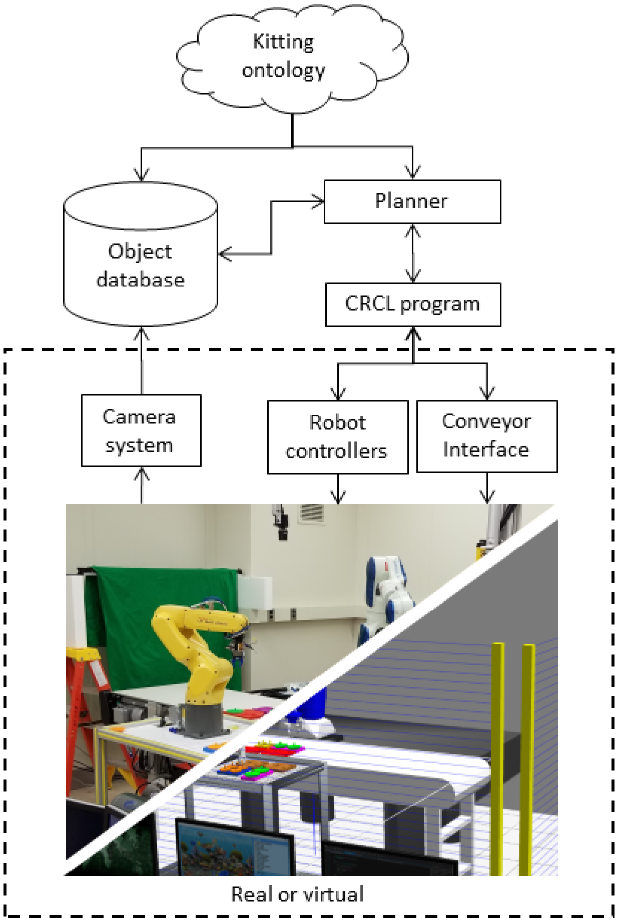


FIGURE 1: APRS system architecture
Components enclosed in dashed box are swappable

Like ROS, Gazebo's functionality can be extended—this time with user-defined plugins, which give access to nearly all aspects of the simulator's physics, sensing, and communication systems. We use plugins to implement both robots' grippers, and the light curtain. The light curtain is implemented as a collection of Gazebo built-in laser rangefinders, whose values are unified into a single output using a plugin. The gripper actuation is handled entirely by a plugin described in the Physics Improvements section. Defining custom behavior using plugins has virtually no performance overhead and can be done in only a few lines of code (as shown in Algorithms 1 and 2).

Our conveyor is implemented using both a ROS controller and a Gazebo plugin. It is modeled as a single planar surface controlled by a linear actuator. The surface needs to apply force to carried objects as though it were moving, while appearing approximately stationary itself (to approximate a realistic conveyor). This is accomplished by using a ROS controller to control the actuator's velocity while using a Gazebo plugin to periodically reset the position of the surface after any small displacement.

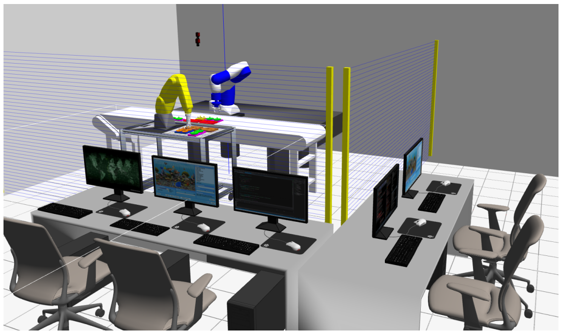


FIGURE 2: The simulated APRS Laboratory

Shared Interface

The physical APRS Laboratory performs low-level control of each robot using proprietary controllers, and operates the conveyor by using a programmable logic controller (PLC) running a Modbus server [22]. These heterogeneous interfaces are unified using the Canonical Robot Command Language (CRCL) [23]. After the APRS architecture synthesizes a high-level plan, this plan is converted into the corresponding CRCL primitives, which are then sent to the appropriate device (either the robot, gripper, or conveyor).

To integrate the simulated robots' controllers with the APRS architecture, we implement a conversion from CRCL primitives to ROS messages, which can be used to specify controller targets and trajectories. The resulting CRCL2ROS library converts CRCL-formatted commands to the ROS-compatible protobuf format [24].

For conveyor commands, we forego using the CRCL2ROS library in favor of re-using the Modbus interface already in-place for the physical conveyor. To do this, we implement a Modbus server in software using the `libmodbus` C++ library [25]. In the physical system, the Modbus server controls the conveyor's velocity by writing to the appropriate registers on the PLC. In the software implementation, the Modbus server controls the simulated conveyor's speed and direction by writing a velocity target to its ROS controller.

4 PHYSICS IMPROVEMENTS

Having established the overall design of the APRS simulation, we will now present two simulation improvements that form the main contribution of this paper. Each improvement targets a component of the simulated system that is prone to unstable, physically unrealistic behavior. “Stability” will be defined and tested more precisely in the “Experiments” section, while this section will focus on the conceptual description of each improvement. We want to stress that these improvements are agnostic to the physics engine used and are simple to implement provided the physics engine supports a programming interface (this is handled by Gazebo’s plugin system in our case).

Parallel Gripper Control

The APRS robots perform grasping using pneumatic parallel grippers. A typical approach for defining parallel grippers in software involves mimic joints, where one joint “mimics” a second joint by maintaining its relative position and velocity in joint-space. Unfortunately, many physics engines do not define a joint constraint that describes this mimic behavior directly. Because of this, a popular ad-hoc approach is to use external control plugins [26]. This approach implements a proportional-integral-derivative (PID) force controller that uses the mimic joint’s relative joint position as the feedback term. However, there are two problems with using such a control scheme for parallel grasping.

- The output force is 0 N when the error feedback is 0 (i.e., when the gripper’s fingers are symmetrically positioned). During grasping, this creates asymmetric forces on the gripper that either cause the grasp to fail or the mimic joint to be pushed to an asymmetric position so that the force generated by the controller matches the force of the original joint.
- Since the original joint’s controller has no knowledge of the mimic joint, there is no guarantee that the gripper’s fingers will remain symmetric throughout a trajectory if they experience different external forces. For example, if an obstacle blocks the mimic joint but not the original joint, the resulting finger positions would not be symmetric like they would be in the real gripper.

To address these shortcomings, we implement a new control scheme operating on both joints simultaneously:

$$F_1 = F_c + k_p(p_2 - p_1), \tag{1}$$

$$F_2 = F_c + k_p(p_1 - p_2). \tag{2}$$

Here, F_i is the force output for each joint; p_i is the joint-space position of each joint; F_c is a constant force shared by both joints; and k_p is a proportional gain. Using this controller, both joints maintain an identical non-zero force during grasps and are driven to symmetric positions by mimicking each other, satisfying both problems.

This control scheme has an elegant physical interpretation: the F_c term emulates the force generated by the gripper’s piston, and the $k_p(\cdot)$ term emulates the normal forces of the gripper’s actuator linkage.

While this controller was developed with pneumatic gripper’s in mind, it can easily be extended for approximating any mimic joint(s). Given some collection of N mimic joints, a single-joint force-controller F_j, and a relative joint-error force-controller F_s, we can define the control of joint $i \in \{1 \ldots N\}$ as:

$$F_i = F_j(p_i) + \sum_{n \in \{1 \ldots N\} \setminus \{i\}} F_s(p_n - p_i). \tag{3}$$

In our gripper implementation, $F_j(\cdot)$ was chosen to be the constant force F_c. $F_s(\cdot)$ was chosen to be a proportional controller.

Contact Simplification

The kitting tasks performed by APRS robots involve handling dozens of small objects such as gears and trays. This introduces two challenges to simulation. The first challenge is simulation stability. The stacking of objects, especially objects with large inertia-ratios, is known to cause instability due to over-constraining the linear complementarity problems (LCPs) that are solved by the physics engine at each time step [5]. The second challenge is compute speed, which deteriorates as the number of contacts to simulate increases.

To improve both stability and computational performance, we implement a Gazebo plugin for automatically simplifying contacts between relatively stationary objects. The approach is straightforward in concept:

1. Use Gazebo’s built-in `ContactSensor` class to find existing contacts and their contact forces
2. Determine whether the objects in contact are stationary relative to each other
3. Create a virtual, fixed joint between objects satisfying the stationary criteria and disable collisions between them
4. Destroy the virtual joint and re-enable collisions if the objects are no longer stationary, resuming normal behavior

Effectively, this replaces a dynamics problem for the physics engine with a much nicer statics problem for the plugin. The sensing of forces, creation and destruction of joints, and enabling and disabling of collisions can be handled entirely through Gazebo’s programming interface. This only leaves the question of how to define the stationary criteria.

For our purposes, two objects are considered stationary if their relative velocity and acceleration are 0. These criteria work fine for creating a virtual joint, but are problematic for deleting it since the virtual joint constrains the relative velocity and acceleration of the objects to be 0. However, we can observe the force needed for the virtual joint to satisfy these constraints and use that as part of the criteria.

More precisely, we can define the net force F_σ on a relatively stationary object as:

$$F_\sigma = F_E + F_C + F_V = 0, \tag{4}$$

where F_E is the external, non-measurable force, F_C is the contact force, and F_V is the virtual joint force. Since only one of either contact forces or joint forces are active at any one time, this equation becomes:

$$F_\sigma = F_E + F_C = 0 \qquad F_\sigma = F_E + F_V = 0. \tag{5}$$

Solving and substituting for F_E, we have $F_C = F_V$ at the timestep that the virtual joint is created. If $F_C \neq F_V$ at a future timestep, then the external force F_E has changed and the original stationary criteria $F_\sigma = F_E + F_C = 0$ is no longer satisfied. The actual plugin uses a threshold on $\|F_C - F_V\|$ rather than a strict inequality, but behaves equivalently.

This plugin offers an improvement over previous dynamics-disabling features found in Gazebo's primary physics engine, which only applied to absolutely stationary objects [5]. It is also fairly physics engine-agnostic and lightweight, provided the physics engine supports measuring constraint forces and dynamically spawning joints. The complete pseudocode is given in Algorithms 1 and 2.

```
Data: GazeboContactSensorMessage Msg,
ErrorThreshold e1, ErrorThreshold e2
Result: VirtualJoint V, ContactForce CF
Executed on new ContactSensorMessage
c1 ← Msg.collision1
c2 ← Msg.collision2
v1 ← GetWorldFrameVelocity(c1)
v2 ← GetWorldFrameVelocity(c2)
a1 ← GetWorldFrameAccel(c1)
a2 ← GetWorldFrameAccel(c2)
if $\|v1 - v2\| < e1$ & $\|a1 - a2\| < e2$ then
    CF ← Msg.NetForce
    V ← GazeboCreateNewJoint(c1, c2)
    GazeboDisableContact(c1, c2)
end
```

Algorithm 1: Virtual Joint Creation Callback

5 EXPERIMENTS

In this section, we measure the performance of our two proposed improvements. To do this, we design several scenarios that target the behavior of each improvement in isolation. Each scenario and metric is designed such that the performance of the real system is known to be trivially simple. For example, we know that once a real robot successfully grasps an object, the object will remain completely stationary relative to the gripper during any of the robot's motions (due to the large force exerted by the pneumatic gripper). Now, the objective is to measure whether the simulation violates this "relatively stationary" condition that is known to hold for the real system. Using this approach, we can verify the physical accuracy of our improvements without needing to take any measurements on the real system.

```
Data: Virtual Joint V, ContactForce CF,
ErrorThreshold e
Executed on each simulation physics step
if $\|CF - V.Force\| > e$ then
    GazeboEnableContact(V.Collision1,
      V.Collision2)
    GazeboRemoveJoint(V)
end
```

Algorithm 2: Virtual Joint Deletion Callback

We tested our scenarios primarily with Gazebo's variant of ODE using default simulation parameters. This configuration was mainly used due to technical limitations in Gazebo's support for other physics engines, and because this ODE configuration is fairly common in other work [1].

Parallel Gripper Control

The gripper controller experiments are motivated primarily by the "relatively stationary" condition described above. We consider a grasp more stable and physically accurate if it maintains a constant displacement between the object and the gripper during a motion. Mathematically, we measure this using the Euclidean distance between an initial position and subsequent positions:

$$\delta_t = \|p_t - p_0\|_2 . \tag{6}$$

Here, p is a 3-dimensional relative position and t is a simulation time-step. This metric is similar to Measure C used in [18].

We would also like to incorporate angular displacement into our analysis. To do this, we record the quaternions q_0 and q_t corresponding to initial and subsequent relative orientations between the gripper and grasped object. Thus, we can compute angular displacement as:

$$\theta_t = \arccos\left(2\langle q_t, q_0\rangle^2 - 1\right). \tag{7}$$

We report this value along with δ_t (ideally both should be close to 0 if the simulation is physically accurate). These metrics are computed for our gripper controller along with two baselines:

	Standard	Plugin
Cumulative δ_t	$1.3 \cdot 10^{-2}$ m	$1.0 \cdot 10^{-3}$ m
Real-Time Factor	0.43	0.98

TABLE 1: Multi-object contact simplification experiment

- Constant opposing forces to both joints
- The preexisting mimic joint plugin from [26] with constant force on the mimicked joint

Each controller is attached to a floating gripper (not experiencing any forces). Using each of the three controllers, the floating gripper grasps a conveniently placed APRS gear. Once the gripper is closed, the floating gripper is subjected to a range of accelerations defined in Cartesian world coordinates (emulating the motion of robots performing pick-and-place operations). During these trajectories, the positional and angular displacement metrics are computed and reported using a combination of ROS and Gazebo telemetry.

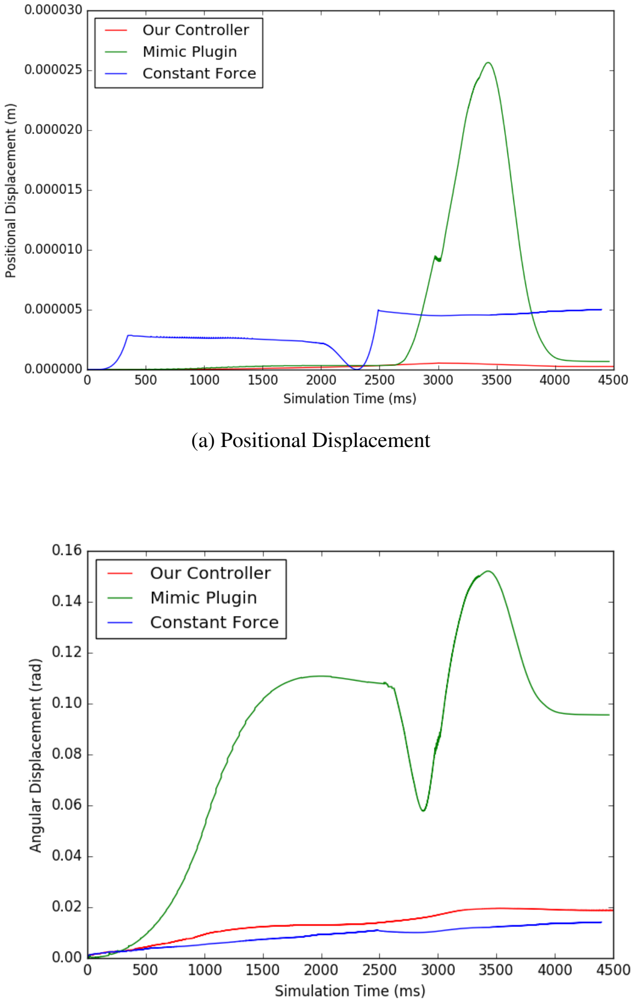
Our results, summarized in Figure 3, show that our proposed parallel gripper controller outperforms both baselines in minimizing positional displacement during Cartesian trajectories. The controller also matches the constant force control scheme in angular displacement. We also note that positional displacement was the largest contributor to objects falling out of the gripper, and that both baselines would drop their objects from much smaller impulses as compared to our proposed controller.

Contact Simplification

Like our previous test, the experiment for our contact simplification improvement uses another “relatively stationary” condition. This time, we know that gears placed inside trays must be stationary relative to the tray (based on their coefficient of friction and the tray design shown in [8]). Therefore, we can reuse the metrics given by Equations 6 and 7, this time computing the relative displacement from gear to tray (rather than gear to gripper).

The scenario that we used consists of a conveyor carrying multiple trays. Each tray is loaded with gears, reminiscent of a typical kitting task. We choose this particular configuration to highlight the benefits of our improvement to both accuracy and run speed for simulations with scaling complexity.

We summarize our results in Table 1. This data was collected based on a simulation of 10 trays carrying 40 gears, which we believe adequately demonstrates the effect of the plugin on both accuracy and simulation speed. We record the relative displacement δ_t between a gear and the tray at $t = 4$ s. We then sum the δ_t values from all 40 gears to compute a cumulative δ_t value for the scenario. We also record the real-time factor of the simulation.



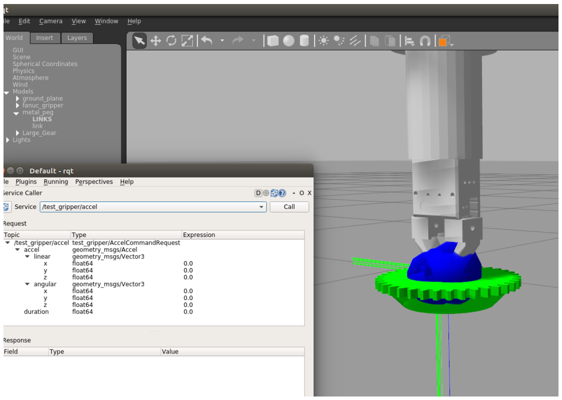
(a) Positional Displacement

(b) Angular Displacement

FIGURE 3: Relative displacement along a Cartesian trajectory using various controllers

As expected, once the contact simplification plugin creates a virtual joint, the displacement of the gear relative to its tray is fixed. This gives the plugin a slight edge over the standard approach by limiting the effect of any unstable simulation behaviour (like jittering contacts).

The plugin also has a pronounced effect on the simulator’s compute speed. On average, every collision between objects in our simulation had about 10 contact points that needed to be processed per time step. The plugin effectively reduced this to one contact point per collision.



Service /test_gripper/accel

Topic	Type	Expression
/test_gripper/accel	test_gripper/AccelCommandRequest	
accel	geometry_msgs/Accel	
linear	geometry_msgs/Vector3	
x	float64	0.0
y	float64	0.0
z	float64	0.0
angular	geometry_msgs/Vector3	
x	float64	0.0
y	float64	0.0
z	float64	0.0
duration	float64	0.0

Field	Type	Value

FIGURE 4: Gripper benchmark tool with Gazebo visualization

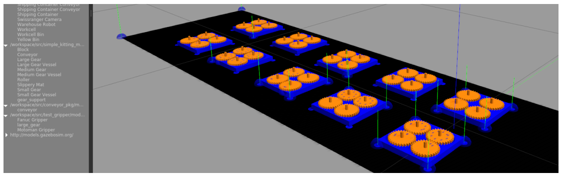


FIGURE 5: Visualization of contact plugin performance testing

6 CONCLUSION

We have presented a physics-based simulation of the APRS Laboratory along with several techniques that enabled it to reach good physical accuracy (as measured by the degree to which the simulation violates static friction conditions). We have demonstrated how our simulation integrates into the overall APRS architecture, and how its design leverages open-source software to include additional industrial components, such as conveyors, grippers, and light curtains.

Given that the APRS Laboratory simulation now achieves good physical accuracy on kitting-related tasks, it can now be extended to simulating entire kitting scenarios currently being developed for the physical APRS Laboratory. To do this, future work will involve improving the software tools needed to create these scenarios, extending the robot agility measures used by the physical lab to the virtual environment, and disseminating the virtual environment to external collaborators. The virtual environment also opens new areas of data-intensive research that are too costly to run on actual hardware, such as controlling robots using reinforcement learning agents or testing the viability of new sensor configurations for computer vision tasks.

Disclaimer

Certain commercial/open source software, hardware, and tools are identified in this paper in order to explain our research. Such identification does not imply recommendation or endorsement by the authors or NIST, nor does it imply that the software tools identified are necessarily the best available for the purpose.

REFERENCES

[1] NIST, 2017. Agile robotics for industrial automation competition. www.nist.gov/el/intelligent-systems-division-73500/agile-robotics-industrial-automation. Accessed 29-April-2019.

[2] Hsu, J. M., and Peters, S. C., 2014. "Extending Open Dynamics Engine for the DARPA Virtual Robotics Challenge". In *Simulation, Modeling, and Programming for Autonomous Robots*, D. Brugali, J. F. Broenink, T. Kroeger, and B. A. MacDonald, eds., Vol. 8810. Springer International Publishing, Cham, pp. 37–48.

[3] Todorov, E., Erez, T., and Tassa, Y., 2012. "MuJoCo: A physics engine for model-based control". In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, pp. 5026–5033.

[4] NVIDIA, 2019. PhysX SDK 4.0. https://developer.nvidia.com/physx-sdk. Accessed: 29-April-2019.

[5] Smith, R., 2006. Open Dynamics Engine. www.gnu-darwin.org/www001/ports-1.5a-CURRENT/devel/ode-devel/work/ode-060223/ode/doc/ode.pdf. Accessed: 29-April-2019.

[6] Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., and Ng, A. Y., 2009. "ROS: an open-source robot operating system". In ICRA workshop on open source software, Vol. 3, Kobe, Japan, p. 5.

[7] OSRF, 2014. Open Source Robotics Foundation - Gazebo. www.gazebosim.org.

[8] Kootbally, Z., Schlenoff, C., Antonishek, B., Proctor, F., Kramer, T., Harrison, W., Downs, A., and Gupta, S., 2018. "Enabling robot agility in manufacturing kitting applications". *Integrated Computer-Aided Engineering,* ***25***(2), pp. 193–212.

[9] Johnson, M., Shrewsbury, B., Bertrand, S., Wu, T., Duran, D., Floyd, M., Abeles, P., Stephen, D., Mertins, N., Lesman, A., Carff, J., Rifenburgh, W., Kaveti, P., Straatman, W., Smith, J., Griffioen, M., Layton, B., de Boer, T., Koolen, T., Neuhaus, P., and Pratt, J., 2015. "Team IHMC's lessons learned from the DARPA robotics challenge trials". *Journal of Field Robotics,* ***32***(2), pp. 192–208.

[10] Aguero, C. E., Koenig, N., Chen, I., Boyer, H., Peters, S., Hsu, J., Gerkey, B., Paepcke, S., Rivero, J. L., Manzo, J., Krotkov, E., and Pratt, G., 2015. "Inside the Virtual Robotics Challenge: Simulating Real-Time Robotic Disaster Response". *IEEE Transactions on Automation Science and Engineering,* ***12***(2), pp. 494–506.

Piliptchak, Pavlo; Aksu, Murat; Proctor, Frederick; Michaloski, John L. "Physics-based Simulation of Agile Robotic Systems." Paper presented at International Mechanical Engineering Congress & Exposition, Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

[11] Bai, Y., Yu, W., and Liu, C. K., 2016. "Dexterous manipulation of cloth". In Proceedings of the 37th Annual Conference of the European Association for Computer Graphics, EG '16, Eurographics Association, pp. 523–532.

[12] Qian, W., Xia, Z., Xiong, J., Gan, Y., Guo, Y., Weng, S., Deng, H., Hu, Y., and Zhang, J., 2014. "Manipulation task simulation using ROS and Gazebo". In 2014 IEEE International Conference on Robotics and Biomimetics (ROBIO 2014), IEEE, pp. 2594–2598.

[13] Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., and Levine, S., 2017. "Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations". *arXiv:1709.10087 [cs]*, Sept. arXiv: 1709.10087.

[14] Erez, T., Tassa, Y., and Todorov, E., 2015. "Simulation tools for model-based robotics: Comparison of Bullet, Havok, MuJoCo, ODE and PhysX". In 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, pp. 4397–4404.

[15] Chung, S.-J., and Pollard, N., 2016. "Predictable behavior during contact simulation: a comparison of selected physics engines". *Computer Animation and Virtual Worlds,* ***27***(3-4), pp. 262–270.

[16] Peters, S., and Hsu, J., 2014. Comparison of Rigid Body Dynamic Simulators for Robotic Simulation in Gazebo. `www.osrfoundation.org/wordpress2/wp-content/uploads/2015/04/roscon2014_scpeters.pdf`. Accessed: 29-April-2019.

[17] Taylor, J. R., Drumwright, E. M., and Hsu, J., 2016. "Analysis of grasping failures in multi-rigid body simulations". In 2016 IEEE International Conference on Simulation, Modeling, and Programming for Autonomous Robots (SIMPAR), IEEE, pp. 295–301.

[18] Kim, J., Iwamoto, K., Kuffner, J. J., Ota, Y., and Pollard, N. S., 2013. "Physically Based Grasp Quality Evaluation Under Pose Uncertainty". *IEEE Transactions on Robotics,* ***29***(6), pp. 1424–1439.

[19] Pepper, C., Balakirsky, S., and Scrapper, C., 2007. "Robot simulation physics validation". In Proceedings of the 2007 Workshop on Performance Metrics for Intelligent Systems, ACM, pp. 97–104.

[20] Carpin, S., Lewis, M., Wang, J., Balakirsky, S., and Scrapper, C., 2007. "USARSim: a robot simulator for research and education". In Proceedings 2007 IEEE International Conference on Robotics and Automation, IEEE, pp. 1400–1405.

[21] Aksu, M., Michaloski, J., and Proctor, F., 2018. "Virtual experimental investigation for industrial robotics in gazebo environment". In Proceedings of the 2018 International Mechanical Engineering Congress and Exposition, ASME.

[22] Swales, A., et al., 1999. "Open MODBUS/TCP specification". *Schneider Electric.*

[23] Proctor, F., Balakirsky, S., Kootbally, Z., Kramer, T., Schlenoff, C., and Shackleford, W., 2016. "The canonical robot command language (CRCL)". *Industrial Robot: An International Journal,* ***43***(5), pp. 495–502.

[24] Google, 2019. Protocol Buffers. `developers.google.com/protocol-buffers/docs/overview?csw=1`. Accessed: 29-April-2019.

[25] Raimbault, S., 2016. libmodbus: A Modbus library for Linux, Mac OS X, FreeBSD, QNX and Windows. `libmodbus.org`.

[26] Chatzilygeroudis, K., 2019. Robotics Group Gazebo Plugins. `https://github.com/roboticsgroup/roboticsgroup_gazebo_plugins/`. Accessed: 3-April-2019.

ON CHARACTERIZING UNCERTAINTY SOURCES IN LASER POWDER BED FUSION ADDITIVE MANUFACTURING MODELS

Tesfaye Moges[1†]
Mechanical Engineering Department,
Indian Institute of Technology Delhi
(IITD), New Delhi, India

Paul Witherell
Engineering Laboratory
National Institute of Standards and
Technology (NIST), Gaithersburg, Maryland

Gaurav Ameta*
Dakota Consulting Inc.
Silver Spring, Maryland

ABSTRACT

Tremendous effort has been dedicated to computational models and simulations of Additive Manufacturing (AM) processes to better understand process complexities and better realize high-quality parts. However, understanding whether a model is an acceptable representation for a given scenario is a difficult proposition. With metals, the laser powder bed fusion (L-PBF) process involves complex physical phenomena such as powder packing, heat transfer, phase transformation, and fluid flow. Models based on these phenomena will possess different degrees of fidelity as they often rely on assumptions that may neglect or simplify process physics, resulting in uncertainty in their prediction accuracy. Predictive uncertainty and its characterization can vary greatly between models. This paper characterizes sources of L-PBF model uncertainty, including those due to modeling assumptions (model form uncertainty), numerical approximation (numerical uncertainty), and model input parameters (input parameter uncertainty) for low and high-fidelity models. The characterization of input uncertainty in terms of probability density function (PDF) and its propagation through L-PBF models, is discussed in detail. The systematic representation of such uncertainty sources is achieved by leveraging the Web Ontology Language (OWL) to capture relevant knowledge used for interoperability and reusability. The topology and mapping of the uncertainty sources establish fundamental requirements for measuring model fidelity and guiding the selection of a model suitable for its intended purpose.

Keywords: Additive Manufacturing, laser powder bed fusion, uncertainty, ontology

1. INTRODUCTION

Additive manufacturing (AM) produces parts by depositing material layer-by-layer without the requirement of specific tooling based on a 3D model [1]. The laser powder bed fusion (L-PBF) process is the most prominent AM technology capable of producing metallic parts with complex geometry and internal structures [2]. It offers the ability to manipulate the chemical compositions and tune part properties by locally controlling the microstructure and mechanical properties to produce finished parts [3]. The process involves different physical activities such as powder layer formation, laser-powder particles interaction, melt pool formation, and solidification. The fundamental physical phenomena that govern the L-PBF process are powder packing, heat transfer, fluid flow, grain growth, and residual stress formation. Due to the complexity of the physical phenomena and process variabilities, the fusion of powder particles affects mechanical properties, surface finish, and fatigue life of the final parts [4].

There are various parameters in the L-PBF process that potentially affect the quality of the part, as well as the energy and material consumption. Some of these parameters include powder size, powder shape, powder size distribution, laser power, spot size, beam profile, layer thickness, scan speed, hatch distance, build orientation, and pre-heat temperature [5]. To insure part quality and ultimately realize the full potential of the L-PBF process, it is crucial to (a) understand the process complexities, (b) identify the sources of process variability, (c) investigate the effect of process parameters, and (d) determine the optimal process parameters. To achieve these goals, intensive experiment-based studies can be time consuming and costly. Thus, efforts have been devoted to computational models and simulations to investigate the parameter-process-structure-property-performance relationships that lead to part quality improvement [6,7]. Although computational models and simulations are beginning to provide a vast resource for predictive models on which design and process decisions can be made, understanding whether or not a model is an acceptable representation for a given scenario is a difficult proposition.

[1] Contact author: tesfaye.moges@nist.gov, tesfaye_mom@yahoo.com

[†] The author is affiliated with IITD, New Delhi, India and the work was done at NIST, MD, USA.

[*] The author contributed to this work while affiliated to Dakota Consulting Inc. Current affiliation is Siemens Corporation, Corporate Technology Office, Princeton, NJ.

Moges, Tesfaye; Witherell, Paul; Ameta, Gaurav. "On Characterizing Uncertainty Sources in Laser Powder Bed Fusion Additive Manufacturing Models." Paper presented at International Mechanical Engineering Congress and Exposition (IMECE2019), Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

Models based on different physical phenomena of the L-PBF process will possess different degrees of fidelity as they often rely on assumptions that may neglect or simplify process physics, resulting in uncertainty in their prediction accuracy. Predictive uncertainty and its characterization can vary greatly between models. For instance, depending on the (a) type of model, such as a physics-based model or an empirical model, or (b) different measurement techniques, such as optical measurements or thermal measurements. Despite these differences, each source of uncertainty can be related back to the fundamental characteristics of the process. Therefore, it is important to investigate and characterize the potential sources of uncertainty in L-PBF models.

Previous work investigated model fidelity, explored assumptions and approximations of various models, and how this information may be captured [5]. Further work investigated specific sources of uncertainty in L-PBF processes, including a thorough review [8]. In this paper, we build on and harmonize the previous works. We characterize sources of L-PBF model uncertainty, including those due to modeling assumptions (*model form uncertainty*), numerical approximation (*numerical uncertainty*), and model input parameters (*parameter uncertainty*) for low and high-fidelity models. First, the L-PBF models are characterized based on their assumptions by capturing the considered and neglected phenomena and input parameters and output quantities of interest (QoIs). These models include the Rosenthal-based thermal models, the finite element method (FEM)-based continuum thermal models, and the powder-scale thermal-fluid flow models [8]. Following this, characterization of the sources of uncertainty is discussed in detail. Model form uncertainty is characterized based on modeling assumptions by capturing the included and neglected physical phenomena during abstraction and formulation of the model. This uncertainty is commonly quantified using a validation metric by comparing simulation results against measurement data. Numerical uncertainty arises primarily from discretization error and is characterized using code and solution verification. Model input parameter uncertainty comes from the inherent variability present in input parameters or those parameters whose exact values are not known and cannot be directly measured. Then, the characterization of models and uncertainty sources is represented using an ontology to capture relevant knowledge that can be used for interpretability and reusability.

The organization of the paper is as follows. We first briefly present the state-of-the-art on uncertainty related studies for L-PBF process in Section 2. We then discuss the characterization of L-PBF models focusing on their assumptions by capturing the considered and neglected phenomena in Section 3. Then, in Section 4, we present the investigation and characterization of L-PBF uncertainty sources. The systematic representation of models and uncertainty sources is described in Section 5. We view this work to be an essential step to establish fundamental requirements for measuring model fidelity, and for guiding the selection of a model suitable for its intended purpose.

2. BACKGROUND

It has been highlighted that computational models and simulations play significant roles in understanding the AM process phenomena and predicting optimum process parameters and hence are important contributing factors to achieving desired part performance and reducing the numbers of faulty parts [4,9]. There has been tremendous effort to develop computational thermal models to understand the thermal history and predict melt pool geometries in L-PBF process. These models can be broadly categorized into three groups: the Rosenthal-based thermal models, the FEM-based thermal models, and the powder-scale thermal-fluid flow models. The Rosenthal-based thermal models analytically solve the heat conduction equation for a moving heat source [10]. The only heat transfer mechanism considered in these models is thermal conduction; those due to convection and radiation are neglected. The models provide low fidelity predictive results, but they are computationally efficient. However, the heat transfer phenomena and phase transformations that exist in the L-PBF process are complex and hence the governing heat transfer equations and boundary conditions that capture these phenomena are difficult to solve analytically. Thus, to understand the thermal history in detail and accurately predict the temperature fields and melt pool characteristics, various numerical models have been developed [8].

The FEM-based thermal models solve the heat transfer governing equations by discretizing the spatial domain into a finite number of elements and the temporal transient phenomena into time steps regardless of geometrical complexity. In these models, the powder bed is considered as a continuum block of material with effective thermo-physical properties, and the heat transfer mechanisms and distribution of absorbed energy are accounted for. However, the phenomena related to melt pool flow are neglected. To understand the fluid flow of the molten pool, more realistic numerical models based on powder-scale, thermal-fluid flow have been developed [11]. These models represent the L-PBF process more realistically by directly accounting for the phenomena related to powder packing (powder size, shape, and size distribution) and melt pool flow (surface tension, shrinkage, recoil pressure, and others). In addition to the common process signatures (temperature fields and melt pool geometries), these models can be used to investigate the formation of different defects such as balling, porosity, and delamination between layers and substrate. The critical challenge of these models is that they are computationally expensive and thus infeasible to use for full part-scale simulations and studying parameter optimizations as these problems require a large number of simulations.

Though experimental-based uncertainty analysis can provide the actual variabilities present in AM processes [12,13], this approach is time consuming and costly. To deploy computational models for design and process decision making and ultimately for part qualification, their degree of fidelity needs to be known first. Thus, understanding the sources of uncertainty is necessary to determine the degree of model fidelity and conduct uncertainty management to identify the main sources of error.

The uncertainty related studies in L-PBF models are relatively new and have started receiving increasing attention in recent years in the AM community [8,14–16]. Some of these studies are reviewed as follows. Moser et al [17] and Ma et al [18] identified the critical input parameters that largely influence the predictive accuracy of the FEM-based thermal model by assigning a probability density function (PDF) to account for parameter variability using a stochastic collocation approach and fractional factorial design of experiment (DOE), respectively. Nath et al [19] conducted uncertainty analysis on the FEM-based thermal model by constructing a statistical surrogate model using a Gaussian process to replace the computationally expensive physics-based thermal model. They also extended the uncertainty analysis to a solidification model to quantify the uncertainty in grain size distribution of microstructure. Kamath [20] conducted uncertainty analysis on the Eagar-Tsai Rosenthal-based thermal model [21] that considers Gaussian-distributed heat source and powder-based thermal model that was developed by Verhaeghe et al [22] using a regression tree and Gaussian process surrogate models. Tapia et al. [23] used polynomial chaos expansions for uncertainty propagation analysis in the Eagar-Tsai Rosenthal-based thermal model and the FEM-based thermal model. Tapia et al. [24] used a Gaussian process surrogate model for the powder-scale thermal-fluid flow model which was developed by Khairallah et al. [25]. Yang et al. [26] investigated different surrogate modeling techniques and discussed the implementation of adaptive sampling method to manage the sources of uncertainty in AM predictive models. There are some research efforts that use deep learning for classification of melt pool size [27] and machine learning for a data-driven continuous construction of AM knowledge [28]. A thorough literature review on machine learning applications in AM can be found in Razvi et al. [29].

The stated previous works are primarily focused on uncertainty analysis related to input parameters to identify the most critical parameters that influence the output QoIs. However, to fully understand the fidelity of a model and perform uncertainty analysis, it is important to investigate all sources of uncertainty: those due to modeling assumptions, numerical approximations, input parameters, and uncertainty due to measurement errors for model validation. Lopez et al [16] made an effort to identify these sources of uncertainty for the Rosenthal-based thermal model considering melt pool width as an output QoI. To quantify uncertainty propagation of input parameters, the Monte Carlo simulation was directly imposed on the physics-based model as the model is computationally efficient. Moges et al [15] extended this work to further investigate and quantify all sources of uncertainty for the Rosenthal-based as well as for the FEM-based thermal models to predict melt pool width. A fractional factorial DOE was used to drive a statistical response surface model on which the Monte Carlo simulation was imposed to quantify parameter uncertainty. Other uncertainty related studies for L-PBF models can also be found in Hu and Mahadevan [30], Mahmoudi et al [31], and Ghosh et al [32]. In present study, we characterize these uncertainty sources and represent them in a systematic fashion to capture their effects on predictive accuracy of computational models.

3. THE L-PBF THERMAL MODELS

The investigation and characterization of sources of uncertainty in computational models begin from understanding the assumptions, abstractions, and governing equations on which these models rely and capturing the included and neglected physical phenomena. Thus, it is important to first characterize the existing thermal models based on their assumptions as it also provides conceptual understanding of model fidelity [5]. In this section, the characterization of computational thermal models based on their assumptions, governing equations, and included and neglected physical phenomena along with input parameters and output QoIs is briefly presented.

3.1 The Rosenthal-based thermal models

The conduction mode of heat transfer is the main governing phenomenon in laser-powder interaction in L-PBF process. The heat conduction equation for a moving heat source is expressed in Equation (1) [33].

$$\rho C_p \frac{\partial T}{\partial t} = (\nabla \cdot k \nabla T) + Q, \tag{1}$$

where ρ is density, C_p is specific heat capacity, k is thermal conductivity, and Q is internal heat. Assuming the heat source distribution as a point or Gaussian heat source moving in the x direction on a surface of a semi-infinite space, Rosenthal [10] and Eagar and Tsai [21] determined the temperature T at a given time t, respectively. The Rosenthal-based thermal models can be used as a foundation to build more realistic approaches by considering the different laser beam distributions (line, cylindrical, Gaussian, or ellipsoid) [33]. The characterization of these models in terms of input parameters, outputs, assumptions, and considered and neglected phenomena is summarized in Table 1.

Table 1: Characterization of the Rosenthal-based thermal models

Input parameters	Laser power, scan speed, absorption coefficient, melting temperature, thermal conductivity, density, specific heat capacity, latent heat of fusion, preheat temperature
Outputs QoIs	Temperature fields, cooling rates, and melt pool dimensions (width and length)
Assumptions	Surface energy distribution, heat source distribution (point, cylindrical, ellipsoid, or Gaussian), continuum material
Considered phenomena	Energy absorptivity, absorbed energy distribution (only cross-section), moving heat source (scan speed), thermal conduction, latent heat of fusion
Neglected phenomena	Absorbed energy distribution (penetration), heat convection, surface radiation, latent heat of vaporization, all phenomena related to melt pool flow (surface tension, Marangoni effect), phase transformation, powder particle packing (powder size distribution, powder particle contact forces)

3.2 The FEM-based thermal models

To capture the steep thermal gradients near the laser spot and heat affected zone and transient nature of the heat transfer phenomena, the FEM-based thermal models are commonly used in L-PBF process [34]. In FEM models, the layer of powder particles is assumed as a continuum block of material with effective thermo-physical properties. The continuum powder bed is discretized into finite elements and the governing heat conduction equation is solved locally with initial condition and boundary conditions. To solve Equation (1) numerically, the initial condition assumed a uniform preheat temperature throughout the powder bed at time $t = 0$ and the boundary conditions on the top surface are given using Equation (2) [34].

$$-k\nabla T \cdot \boldsymbol{n} = q + h(T - T_0) + \varepsilon\sigma(T^4 - T_0^4), \tag{2}$$

where $\boldsymbol{n}$ is the vector normal to the surface, q is thermal heat flux, h is convection coefficient, ε is thermal radiation coefficient, T_0 is preheat temperature, and σ is Stefan-Boltzmann constant. The surface heat flux of the laser beam is given using Equation (3) assuming Gaussian heat source distribution [35].

$$q = \frac{2AP}{\pi r_b^2} \exp\left(\frac{-2r^2}{r_b^2}\right), \tag{3}$$

where A is absorption coefficient, P is laser power, r_b is laser spot radius, and r is radial distance.

Similarly, the characterization of the FEM-based thermal models in terms of input parameters, outputs, assumptions, and considered and neglected phenomena is summarized in Table 2.

Table 2: Characterization of the FEM-based thermal models

Input parameters	Laser power, scan speed, absorption coefficient, latent heat of fusion, solidus temperature, liquidus temperature, thermal conductivity, density, specific heat capacity, preheat temperature, layer thickness, beam radius, emissivity, convection coefficient
Outputs QoIs	Temperature field, melt pool dimensions (width, depth, and length), melt pool shape
Assumptions	heat source distribution (line, double ellipsoid, or Gaussian), continuum powder bed
Considered phenomena	Energy absorptivity, moving heat source (scan speed), absorbed energy distribution (cross section and penetration), thermal conduction, heat convection, surface radiation, latent heat of fusion
Neglected phenomena	Latent heat of vaporization, all phenomena related to melt pool flow (surface tension, Marangoni effect, etc.), phase transformation, powder particle packing (powder size distribution, powder particle contact forces)

3.3 The powder-scale thermal-fluid flow models

Further understanding of the L-PBF process can be achieved by accounting for the actual physics of the process. As stated above, these include considering the powder bed as distributed powder particles instead of a continuum block of matter and incorporating the effect of melt pool flow as well as the gas-liquid-solid interactions. Powder-scale thermal-fluid flow models are able to consider these physical phenomena and simulate the L-PBF process. In these models, different physical phenomena that potentially govern the fusion process and part quality can be captured. To simulate these physical phenomena, the 3D transient conservation equations of mass continuity, momentum, and energy are solved numerically. The conservation equations are expressed using Equations (4)-(6) [36].

$$\nabla \cdot (\rho\vec{v}) = 0\,, \tag{4}$$

$$\frac{\partial}{\partial t}(\rho\vec{v}) + \nabla \cdot (\rho\vec{v} \otimes \vec{v}) = \nabla \cdot (\mu\nabla\vec{v}) - \nabla \mathrm{p} + \rho\vec{\mathrm{g}} + F_b, \tag{5}$$

$$\frac{\partial}{\partial t}(\rho h) + \nabla \cdot (\rho\vec{v}h) = q + \nabla \cdot (k\nabla T), \tag{6}$$

where $\vec{v}$ is velocity vector, μ is viscosity, p is pressure, $\vec{\mathrm{g}}$ is gravitational acceleration vector, F_b is buoyancy force, h is enthalpy, and $\otimes$ is the Kronecker product. The position of the free surface in terms of phase fraction (F) at the void-liquid interface as a function of time is tracked using the volume of fluid (VOF) method and it is expressed in Equation (7) [37].

$$\frac{\partial F}{\partial t} + \nabla \cdot (\vec{v}F) = 0, \tag{7}$$

The Equations (4) to (7) are solved together to provide the 3D transient temperature and velocity fields using boundary conditions such as the heat exchange between the top surface and the surroundings and the Marangoni shear stress induced by the special variation of surface tension which are expressed using Equations (8) and (9) [37].

$$-k\nabla T \cdot \boldsymbol{n} = h(T - T_0) + \varepsilon\sigma(T^4 - T_0^4) + q_{evap}, \tag{8}$$

$$\gamma(T) = \gamma_m + \frac{d\gamma}{dT}(T - T_m), \tag{9}$$

where q_{evap} is evaporation heat loss, γ is the surface tension at the surface temperature T, γ_m is the surface tension at the melting temperature, and $\frac{d\gamma}{dT}$ is the temperature coefficient of surface tension.

In addition to the convective and radiative heat loss from the top free surface, under intense laser irradiation, heat loss due to evaporation and recoil pressure need to be considered [37]. To predict the amount of energy absorbed by the powder bed, a 3D volumetric heat source that accounts for multiple reflections of the laser beam inside the powder layers has to be considered [7]. Moreover, the temperature dependent material properties of the powder material significantly affect the accuracy of the L-PBF models. Since the powder bed consists of powder particles as well as shielding gas within the interparticle space, the effective temperature dependent properties, such as density, specific heat, and thermal conductivity of the powder bed depend on properties of the powder material and shielding gas [38]. The packing structure of the powder bed in terms of packing density plays a crucial role in simulating the L-PBF process and needs to be coupled with the thermal-fluid flow models. The choice of powder particles distribution (Gaussian, bimodal, uniform, or mono-sized) significantly influences the packing structure of the powder bed in terms of packing density and porosity. The powder bed models are predominantly characterized by particle

Table 3: Characterization of powder-scale thermal-fluid flow models

Input parameters	Laser power, scan speed, absorption coefficient, latent heat of fusion, solidus temperature, liquidus temperature, boiling temperature, thermal conductivity, density, specific heat capacity, ambient temperature, layer thickness, beam radius, emissivity, convection coefficient, surface tension coefficient, viscosity, latent heat of evaporation
Outputs QoIs	3D transient temperature fields and velocity distributions, melt pool dimensions (width, depth, and length), melt pool shape, surface roughness, geometric dimension, porosity, voids
Assumptions	heat source (Gaussian or double ellipsoid), Newtonian and incompressible molten metal
Considered phenomena	Energy absorptivity, absorbed energy distribution (cross section and penetration), multiple reflections effect, thermal conduction, heat convection, surface radiation, latent heat of fusion and evaporation, melt pool flow (surface tension, Marangoni effect, buoyancy, gravity, recoil pressure, mass change due to evaporation and condensation), phase transformation (melting, solidification, vaporization, condensation), powder particle packing (powder size distribution, powder particle contact forces: collision, friction, adhesion)
Neglected phenomena	Mass change due to chemical reaction (oxidation), solid-state phase transformation, gas flow and interaction with solid and liquid, chemical element diffusion and chemical reaction.

shape and size, particle size distribution, packing density, layer thickness, and re-coater velocity and geometry [8]. The heat transfer-fluid flow models can also be coupled with the solidification and residual stress and distortion models to determine microstructure and solidification parameters and mechanical properties of the fabricated parts [39]. The characterization of the powder-scale thermal-fluid flow models in terms of input parameters, outputs, assumptions, and considered and neglected phenomena is summarized in Table 3.

4. CHARACTERIZING L-PBF UNCERTAINTY SOURCES

The beginning of verification, validation, and uncertainty quantification (V&V UQ) for any scientific computing is to identify and characterize the sources of uncertainty [40]. The flow of uncertainty sources and the verification and validation and UQ adapted from Assouroko et al [5], Lopez et al [16], and ASME V&V-20 Standard [41] for L-PBF computational models is depicted in Figure 1. Using the basic principles of physical laws, such as conservation of mass, momentum, and energy, mathematical models are abstracted and formulated based on assumptions to represent the L-PBF physical process. Such assumptions during mathematical model development cause model form uncertainty that results in inaccurate prediction of output QoIs. To solve the partial differential equations and simulate the L-PBF physical phenomena, computational models use numerical approximations. Such approximations cause numerical uncertainty that undermines the predictive accuracy of the models. The V&V UQ process involves verification that verifies whether the computational model accurately solves the mathematical equations and quantifies the numerical uncertainty due to discretization errors. The validation process evaluates how accurately the computational model represents the L-PBF physical process and estimates the model bias and model form uncertainty by comparing measurement results along with associated uncertainty against simulation results along with numerical and parameter uncertainties. The parameter uncertainty of an output QoI is estimated by propagating uncertainties of input parameters through computational model or surrogate model (if the model is computationally expensive). This section discusses the characterization of these sources of uncertainty through V&V UQ approaches.

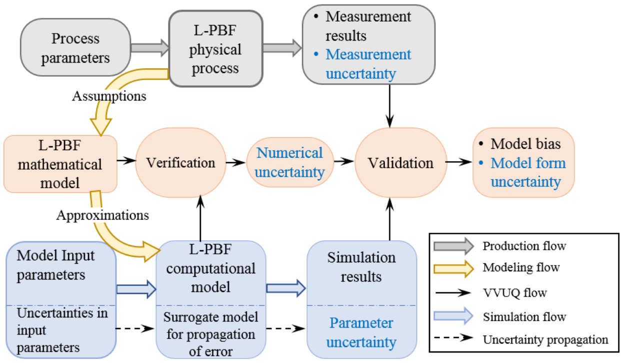


FIGURE 1: UNCERTAINTY SOURCES AND V&V UQ IN L-PBF COMPUTATIONAL MODELS

4.1 Model form uncertainty

The mathematical models only capture certain physical phenomena of the L-PBF process as they are abstracted and formulated based on assumptions and simplifications, and hence they cannot exactly represent the physical mechanisms of the process. As discussed in Section 3, there is a wide range of L-PBF thermal models ranging from low to high fidelity and models within the same level of fidelity. The predictive accuracy of these models potentially depends on the assumptions, considered, and neglected process physics (Tables 1-3). Model form uncertainty arises due to assumptions associated with physical phenomena that neglect or simplify some physics of the process. The predictive accuracy of models across different fidelity or within the same level of fidelity is different due to model form uncertainty. For instance, the assumptions and simplifications of physical phenomena associated with boundary conditions and temperature dependent properties in modeling of heat transfer and phase transformations can lead to inaccurate prediction of temperature gradient and melt pool geometry [8]. The assumptions associated with the distribution of heat source as a point, line, or double ellipsoid heat source significantly conflict with the measured power density distribution. Moreover, ignoring the convective heat transfer, which is one of the main mechanism of heat transfer within the melt pool, can lead to highly inaccurate temperature fields and cooling rates [7].

Model form uncertainty is commonly characterized using a validation approach that evaluates the predictive accuracy of a model by comparing simulation results against measurement results collected under the same condition [42]. Since uncertainties associated with model input parameters propagate through the model and potentially affect the accuracy of simulation results, the values of these parameters need to be first measured and their uncertainties need be characterized. Second, the measurement results along with their uncertainties need to be obtained to perform an effective validation process and estimate model form uncertainty. The ASME V&V 20 Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer [41] discussed the verification and validation activities, along with quantifying sources of uncertainty, in heat transfer and computational fluid dynamics. Since L-PBF process involves heat transfer and melt pool flow phenomena, this standard can be suitable for characterizing sources of uncertainty in L-PBF models. The interval within which model bias falls is characterized by validation metrics and expressed using the following expression (10).

$$\delta_{model} \in [E - u_{val}, E + u_{val}], \tag{10}$$

where E is the comparison error between simulation result and measurement result and u_{val} is validation uncertainty. The validation uncertainty that accounts for the sources of uncertainty due to model input parameters (u_{input}), numerical approximations (u_{num}), and measurement errors (u_D) is evaluated using Equation (11). Therefore, all sources of uncertainty need to be characterized for proper assessment of predictive accuracy of L-PBF models.

$$u_{val} = \sqrt{u_{num}^2 + u_{input}^2 + u_D^2}, \tag{11}$$

4.2 Numerical uncertainty

The governing mathematical equations that capture the complex L-PBF physical phenomena are difficult to solve analytically. Thus, numerical methods are often used to solve these equations based on simplification and approximation to obtain approximate solutions. This approximation introduces numerical uncertainty into the predicted QoIs. There are different sources of uncertainty associated with numerical approximations in any computational simulations, such as truncation error, discretization error, error due to computer programming mistakes, iteration error, and round-off error [43]. Among these sources of uncertainty, discretization error is often considered as the main source of uncertainty due to its large magnitude and it is the center of attention in most verification related studies [43]. The verification process insures that a numerical model accurately represents the underlying mathematical equations and estimates numerical uncertainty due to discretization errors. Thus, numerical uncertainty associated with numerical approximations is characterized by the model verification process. To address this matter, verification is divided into two fundamental parts: code and solution verifications [44].

Code verification makes sure that an algorithm or a computer code is free of mistakes or bugs. This is achieved using software quality assurance (SQA) techniques or the method of manufactured solutions (MMS) by comparing predictive outputs with analytical solutions or manufactured solutions if analytical solutions are difficult to compute [43]. In L-PBF models, different simulation codes have been used to numerically estimate the output QoIs. These include discrete element method (DEM) codes to determine powder packing density, OpenFOAM, Flow-3D, or ALE3D code to simulate heat transfer and fluid flow as well as other commercial codes and softwares like ANSYS and ABAQUS to predict the QoIs in heat transfer and residual stress analyses [8]. The Rosenthal-based analytical solution of temperature field is commonly utilized to conduct code verification for L-PBF thermal models. Hence, code verification needs to be conducted to make sure that these codes have satisfied the order of accuracy test.

Solution verification estimates the sources of uncertainty associated with numerical approximations. The governing partial differential equations are solved by discretizing the spatial domain of interest into a finite number of elements and the time advancement into a finite time step. This approach is common in L-PBF models that use different numerical methods, such as FEM, FDM, FVM, DEM, LBM, and CFD [8]. This discretization causes numerical uncertainty in computational simulations. There are different techniques to quantify this source of uncertainty, such as Richardson extrapolation, Roache's grid convergence index, and others [43]. Discretization error is commonly estimated by comparing predictive outputs determined at different grids with course, medium and fine mesh sizes. Roache [45] used grid convergence index (GCI) to convert the discretization error determined from Richardson extrapolation into uncertainty, and the equations used to compute numerical uncertainty are given in Equations (12)-(14). A detailed explanation and formulation on code and solution verifications taking heat transfer and computational fluid dynamics models as a case study can be found in ASME V&V 20 Standard [41] and Roy and Oberkampf [43]. Since numerical-based L-PBF models are computationally expensive, statistically-driven surrogate models are commonly developed to overcome this issue. This introduces surrogate model uncertainty due to the limited number of training and testing data.

$$GCI = F_s \frac{\epsilon_{ext}}{r_{21}^p - 1}, \tag{12}$$

$$\epsilon_{ext} = \left|\frac{f_{\text{ext}} - f_1}{f_{\text{ext}}}\right|, f_{\text{ext}} = \frac{r_{21}^p f_1 - f_2}{r_{21}^p - 1}, p = \frac{\ln(|f_3 - f_2| / |f_2 - f_1|) + q(p)}{ln r_{21}}, \tag{13}$$

$$\begin{gathered} q(p) = ln\left[\left(r_{21}^p - s\right)/\left(r_{32}^p - s\right)\right], \\ s = sign[(f_3 - f_2)/(f_2 - f_1)], \end{gathered} \tag{14}$$

where GCI is grid convergence index, F_s is factor of safety, f_1, f_2, and f_3 are simulation results at course h_1, medium h_2, and fine h_3 mesh sizes, r_{21} and r_{32} are the mesh refinement ratios, p is order of convergence. For a constant mesh refinement ratio, $q(p) = 0$. Otherwise, the order of convergence is computed recursively with initial guess $q(p) = 0$. A numerical

Moges, Tesfaye; Witherell, Paul; Ameta, Gaurav. "On Characterizing Uncertainty Sources in Laser Powder Bed Fusion Additive Manufacturing Models." Paper presented at International Mechanical Engineering Congress and Exposition (IMECE2019), Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

uncertainty is computed by multiplying the GCI with a simulation result.

4.3 Parameter uncertainty

Computational models utilize different input parameters to simulate the behavior of physical phenomena existing in L-PBF process. There is inherent variation in process parameters and in some cases, precise values of some parameters are not known due to imprecise measurement methods or inaccurate estimation of physical properties as the L-PBF process exhibits phase transformation at a small length scale within a very short period of time. Thus, the sources of parameter uncertainty come from natural variability existing in process parameters and/or due to parameters whose exact values are not known or cannot be directly measured. For instance, temperature dependent material properties possess uncertainty due to (a) difficulty in accurately measuring their values especially at high temperature, (b) availability and usage of different measuring techniques, or (c) different values reported in the literature [8]. As a result, any uncertainty associated with input parameters propagates into simulation outputs QoIs through computational models and results in reduction of predictive accuracy. Depending on the amount of information known regarding the distributions of parameter uncertainty, the model input parameter uncertainty can be commonly classified into two categories: aleatory and epistemic uncertainty [14].

Aleatory uncertainty is described as the uncertainty that arises due to natural variation or randomness in a system. There are a significant number of sources of parameter uncertainty in L-PBF process that fall under this category. These include uncertainty sources due to variation in powder size, shape, and size distribution, fluctuations and inherent drift in laser system and galvanometer, vibration in motion and position of built platform and re-coater arm that alter layer thickness, and others. The uncertainty sources that cause variation in process parameters, temperature dependent properties, and absorption coefficient are discussed in Moges et al [8]. This type of uncertainty is characterized by a distribution function, either a probability density function (PDF) or a cumulative distribution function (CDF) to represent the frequency of occurrence [42]. To define the distribution of parameter uncertainty, enough data/knowledge is required.

Epistemic uncertainty is a type of uncertainty that arises due to lack of knowledge and thus it can be reduced by introducing additional information. If a distribution function is assumed for input parameter uncertainty without having enough information, it introduces additional uncertainty into output QoIs. Such an assumption is common in L-PBF input parameters due to the limited number of samples to precisely define the form and parameters of the distribution function. For example, there is limited and sparse data of absorption coefficient and coefficient of friction parameters to accurately define the type and parameters of distribution functions. This uncertainty is commonly characterized by a distribution function to represent the degree of belief [42].

There are different approaches to propagate sources of parameter uncertainty through a model and quantify uncertainty of output QoIs. These include Monte Carlo sampling, response surface methods, stochastic collocation method, polynomial chaos expansion, Gaussian process model, support vector machines, and others [42,46]. For instance, Monte Carlo sampling randomly select a number between zero and one and applies a distribution function to obtain the corresponding parameter sample. This method requires a large number of simulations and hence is only applicable for models that are computationally efficient. For intensive simulation models, the common approach for propagating input parameter uncertainty sources is through surrogate models, such as polynomial chaos expansion, Gaussian process, and others. For example, polynomial chaos expansion expresses output QoI in terms of model input parameters along with their uncertainties using orthogonal basis functions and coefficients and the general expression is given by Equation (15) [14,23].

$$y(\mathbf{x}) = \sum_{j=0}^{N_b} a_j \psi_j(\mathbf{x}), \tag{15}$$

where $\mathbf{x}$ is set of input variables, $y(\mathbf{x})$ is output response, a_j is coefficient of basis function, $\psi_j(\mathbf{x})$ is orthogonal basis function, and N_b is number of basis functions.

4.4 Measurement uncertainty

To accurately compare simulation results with experimental data and perform a validation process to estimate model form uncertainty, measurement results need to be provided along with associated uncertainty of the output QoIs. Measurement uncertainty originates from imprecise measurement methods and error in equipment calibration. Different measurement techniques, such as optical measurements or thermal measurements to conduct in-process monitoring, result in different uncertainty in surface temperature due to the difference in measurement methods [47]. Measurement uncertainty is mainly characterized by different components, such as mean, standard deviation, and probability distribution. These components can be evaluated using statistical method by utilizing the results from a series of measurements. The standard way to evaluate measurement uncertainty from experimental data is described in the “Guide to the Expression of Uncertainty in Measurement (GUM)” [48]. The confident interval of the measurement uncertainty is derived from the standard deviation of a sample population which is evaluated using Equation (16).

$$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{1/2}, \tag{16}$$

where s is standard deviation, n is number of measurements, y_i is measured QoI, and $\bar{y}$ is mean value of all measurement results.

5. CASE STUDY: CHARACTERIZING UNCERTAINTY SOURCES IN L-PBF THERMAL MODEL

In this case study, a Rosenthal-based semi-analytical melt pool model is selected to demonstrate the concept presented in section 4. This semi-analytical melt pool model solves the heat conduction equation for a moving heat source (Equation 1) using a set of ordinary differential equations that describes the motion of isotherms on the surface of the powder bed. By assigning one of the isotherms to the melting temperature, the model can be used to predict the melt pool width. This model is first developed for laser cladding using isotherm migration method [49] and adjusted for use in L-PBF process [16]. The method extends the Rosenthal's solution for moving heat source to consider the temperature-dependent material properties. This model is highly simplified thermal model that neglects multiple physical phenomena of the L-PBF (Table 1), it is suitable for quick prediction of melt pool dimensions.

All sources of uncertainty described in section 4 exist in this semi-analytical melt pool model. Due to the assumptions in terms of point heat source distribution, continuum powder bed, and ignoring melt pool flow phenomena (see Table 1), the model comprises of model form uncertainty associated with these assumptions. The temperature increment used to assign the isotherms on the powder bed surface induces discretization error that causes numerical uncertainty. The uncertainty associated with input parameters propagates through the model and causes uncertainty on the predicted melt pool width. Measurement uncertainty associated with melt pool width is also plays major role in model validation. The detailed analysis and UQ on this regard is given in Moges at al [15].

As discussed in section 4, estimation of modeling error begins from code and solution verification. It was reported that the code verification for the model used in this case study was conducted using manufactured solution and the method converges to the Rosenthal's analytical solution [16]. The solution verification that estimates numerical uncertainty due to discretization error is conducted using Roache's grid convergence index. The GCI that estimates the 95% convergence is determined using Equation (12) based on the values of melt pool width obtained at course, medium, and fine grid refinements. The estimated melt pool width along with numerical uncertainty at laser power 195 W, scan speed 800 mm/s, and absorption coefficient 0.6 for IN625 material is (127.5 ± 2.3) µm.

The source of uncertainty due to unknown input parameters is quantified using full factorial design of experiment (DOE) analysis assuming normal distribution for the parameters given in Table 4. A statistically-driven surface response that uses to estimate parameter uncertainty of the melt pool width is first derived from DOE analysis. Then Monte Carlo simulation is conducted to determine the probability distribution of melt pool width. The histogram of the melt pool widths for 10,000 samples and fitted normal distribution is shown in Figure 2. The average of the predicted melt pool widths and the parameter uncertainty that represents the 95% confidence interval is obtained to be (138.7 ± 12.8) µm.

Table 4: Input parameters for uncertainty propagation

Input parameters	Nominal value	% variation
Laser power	195 W	2.5%
Scan speed	800 mm/s	1.5%
Absorption Coefficient	0.6	20%
Heat capacity	$c_p(T)\ J/kgK$	3%
Thermal conductivity	$k(T)\ W/mK$	3%

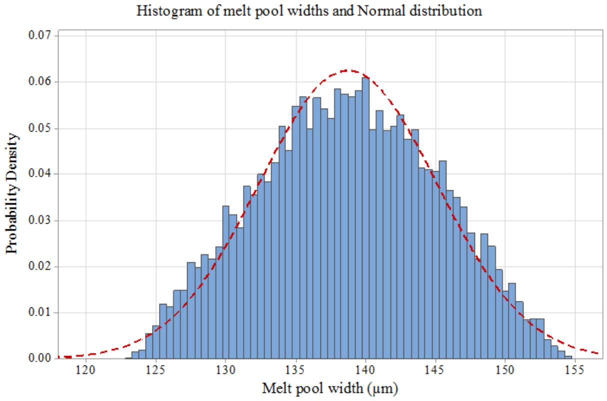


FIGURE 2: PROBABILITY DENSITY OF PREDICTED MELT POOL WIDTHS

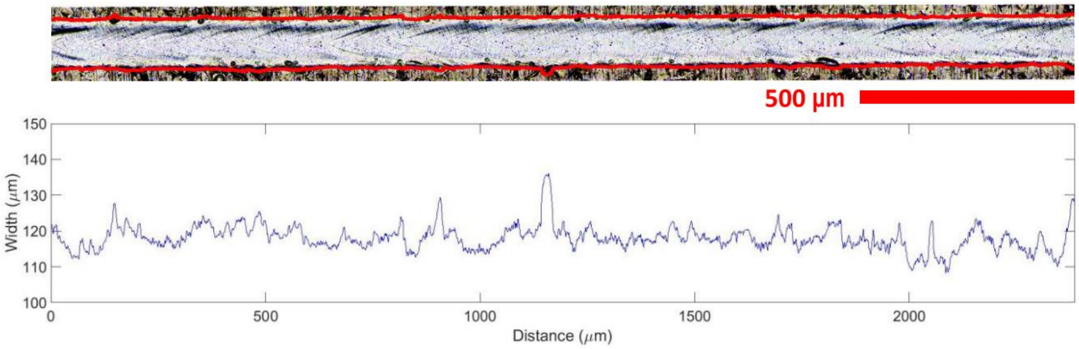


FIGURE 3: MEASUREMENT OF MELT POOL WIDTH FROM OPTICAL IMAGE OF SCAN TRACK. FIGURE ADOPTED FROM FOX ET AL. [50]

To determine measurement uncertainty for model validation and estimate model uncertainty, melt pool width was measured from the image of scan track taken using optical microscope by manually tracing the edges of the scan track and using a software to determine the distance between the traces as shown in Figure 3 [50]. The melt pool average and standard deviation at 195 W laser power and 800 mm/s scan speed are 132.2 μm and 14.1 μm, respectively. Thus, the 95% confidence interval is ± 28.2 μm. In addition, manually tracing the edges of the track induces ± 2 μm uncertainty.

The validation uncertainty, assuming all uncertainty sources are independent, is determined using Equation (11) and is estimated to be (127.5 ± 32.9) μm. The comparison between the predicted and measured melt pool width and the confidence interval obtained from validation uncertainty is depicted in Figure 4.

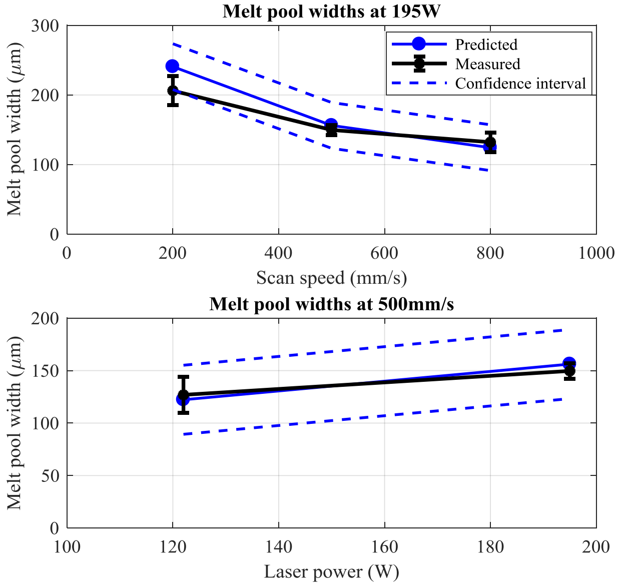


FIGURE 4: PREDICTED AND MEASURED MELT POOL WIDTH AND 95% CONFIDENCE INTERVAL

6. REPRESENTATION OF L-PBF MODELS AND UNCERTAINTY SOURCES

To capture relevant knowledge of the L-PBF models and sources of uncertainty, systematic representation of their characterization is essential. For this purpose, the Web Ontology Language (OWL)-based ontology is leveraged to extract relevant knowledge that can be useful for interoperability and reusability. There are some research efforts that use ontology-based knowledge representation in AM [5,51,52]. This section presents the addition of specific classes and attributes into our previous AM ontology [5] in order to capture knowledge associated to particle-scale thermal-fluid flow models and sources of uncertainty. First, we discuss the representation of the L-PBF models focusing on the powder particle-based models, and then we present the representation of uncertainty sources focusing on their classifications and their relationships with the models.

6.1 Representation of L-PBF models

The proposed ontology in the present study captures the formulations, assumptions, input parameters, output QoIs, and predictive models of the five main physical mechanisms of the L-PBF process: powder layer deposition, heat source-powder interaction, melt pool formation, solidification and grain growth, and the occurrence of residual stress and distortion formation. The main class named *LPBFModel* comprises of the following subclasses: *Formulation*, *Assumption*, *InputParameters*, and *Prediction,* with prefix of *LPBFModel,* as well as *LPBFPredictiveModel*. The *LPBFModelFormulation* subclass involves the mathematical formulations used to capture the aforementioned physical mechanisms of the L-PBF process. For instance, the *MeltPoolModelFormulation* includes the conservation equations of mass, momentum, and energy that capture the heat transfer and fluid flow phenomena in the melt pool. The modeling assumptions and simplifications used while formulating the physical phenomena are captured under the subclass called *LPBFModelAssumption*. For example, this subclass captures the assumptions associated to distribution of heat source: point, cylindrical, ellipsoidal, or Gaussian; powder bed material distribution: continuum or distributed powder particle; powder size distribution: mono-size, bi-modal, uniform, Gaussian, or positively skewed. The subclass that captures the input parameters including process parameters: laser power, scan speed, layer thickness, beam size, etc. and material properties: thermal conductivity, density, heat capacity, melting temperature, etc., is *LPBFModelInputParameter*.

The output QoIs of the L-PBF predictive models are captured by the *LPBFModelPrediction* subclass. This involves the outputs of (a) powder bed model: packing density, coordination number, and radial distribution function; (b) heat source model: absorbed energy, absorbed energy distribution, and effective absorption coefficient; (c) melt pool model: temperature gradient, melt pool dimensions: width, depth, and length, defects: balling, keyhole, and layer delamination, and porosity: gas pores and inter-track voids; (d) solidification model: grain size, grain morphology, and grain orientation; (e) residual stress and distortion model: residual stress and strain distribution, deformation history, fatigue life, and shrinkage. The last subclass under *LPBFModel* is *LPBFPredictiveModel* and this captures the different predictive models in L-PBF process. These models include (1) powder bed models: raindrop method and discrete element method (DEM); (2) heat source model: Beer Lambert, ray tracing, radiation transfer, and surface heat flux; (3) melt pool models: Rosenthal-based model, FEM thermal model, CFD-based model, and lattice Boltzmann method; (4) solidification model: phase field method and cellular automata; (5) residual stress and distortion model: thermomechanical FEM-based model and simplified mathematical model.

6.2 Representation of L-PBF uncertainty sources

The uncertainty related aspects of the L-PBF models, which include sources of uncertainty, uncertainty quantification approaches, and types of uncertainty, are captured in the proposed ontology. The *LPBFUncertainty* class involves subclasses such as *TypeOfUncertainty*, *LPBFUncertaintySource*, and *LPBFUQMethod*. As discussed in previous section, the *TypeOfUncertainty* in any scientific computing can be categorized as *AleatoryUncertainty*, *EpistemicUncertainty*, or *CombinedUncertainty*. As mention in previous section, aleatory uncertainty is due to natural randomness of input quantities or perturbation in a system and cannot be reduced. There are different uncertainty sources that fall under this category. These include variation in measurement results, inherent drift in laser supply system, fluctuation in laser power, scan speed, and beam radius, variation in powder size, shape, and size distribution, friction coefficient, and absorption coefficient of a powder bed. On the other hand, epistemic uncertainty comes from lack of knowledge and can be reduced by introducing more information. The L-PBF uncertainty sources that fall under this category include uncertainty due to modeling assumptions that neglect some physical phenomena (model form uncertainty), uncertainty due to numerical approximation (numerical uncertainty), limited measurement data, error in instrument calibration, imprecise measurement method, and uncertainty in distribution type and parameters due to sparse data.

The *LPBFUQMethod* represents different methods and standard approaches to quantify model form, numerical, parameter, and measurement uncertainties. For instance, measurement uncertainty can be quantified using a standard approach GUM, numerical uncertainty by verification approach, model form uncertainty by validation approach, and parameter uncertainty by sampling methods or surrogate models. The possible sources of uncertainty in L-PBF models including those due to measurement error are identified and captured under the subclass called *LPBFUncertaintySource.* The taxonomy of the uncertainty sources of the top-level entities is depicted in Figure 5.

6.3 Relationships in L-PBF models and uncertainty sources

To capture the interactions between the features of L-PBF phenomena, parameters, and output QoIs at the physical and computational domains, properties that define the relationships are established in the proposed ontology. As described in section 3, the L-PBF models are developed based on assumptions and they do not consider the entire phenomena of the process. Thus, the properties that link these models with their corresponding inputs, outputs, assumptions, and the captured and neglected phenomena are defined in the ontology using *requires*, *predicts*, *assumes*, *considers*, and *neglects*, respectively. Figure 6(a)-(c) shows the relationships in semi-analytical Rosenthal-based thermal model, FEM-based thermal model, and powder-scale thermal-fluid flow model. Similarly, the sources of uncertainty that result in model discrepancy are defined by properties to link different sources with corresponding predictive uncertainties. Those features that cause model form, numerical, parameter, and measurement uncertainties are being captured in the ontology as depicted in Figure 7 and the domains and ranges of properties are clearly defined. For instance, model form uncertainty in (a) powder bed model is caused by assumptions associated with powder bed material distribution and powder size distribution; (b) heat source model is caused by dimensionality of absorbed energy (surface and/or volumetric) and distribution of heat source, and (c) melt pool model is caused by thermal boundary conditions, initial conditions, phase transformation assumptions, and molten metal flow assumptions.

The features that cause variability in some input parameters are captured under parameter uncertainty sources. For example, variability in (a) laser power can be caused by heating of optics, soot on optics and inherent drift in laser delivery system and (b) layer thickness is caused by orientation and positioning errors, vibration in build platform motion, vibration in recoater arm motion, and variation in powder bed density. The main source of numerical uncertainty is discretization errors due to the selection of element size and time steps. Lastly, measurement uncertainty is caused by calibration error, imprecise measurement methods, and variation in measurement results.

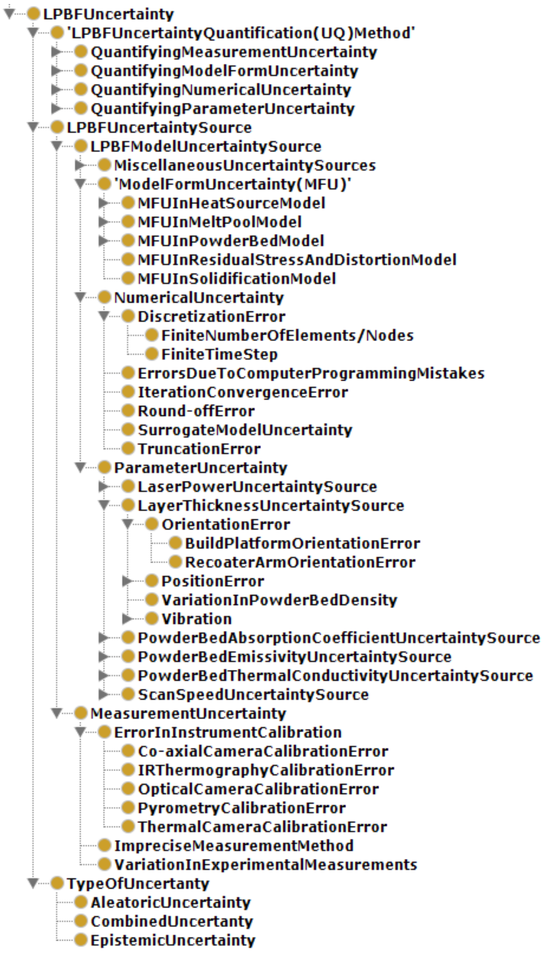
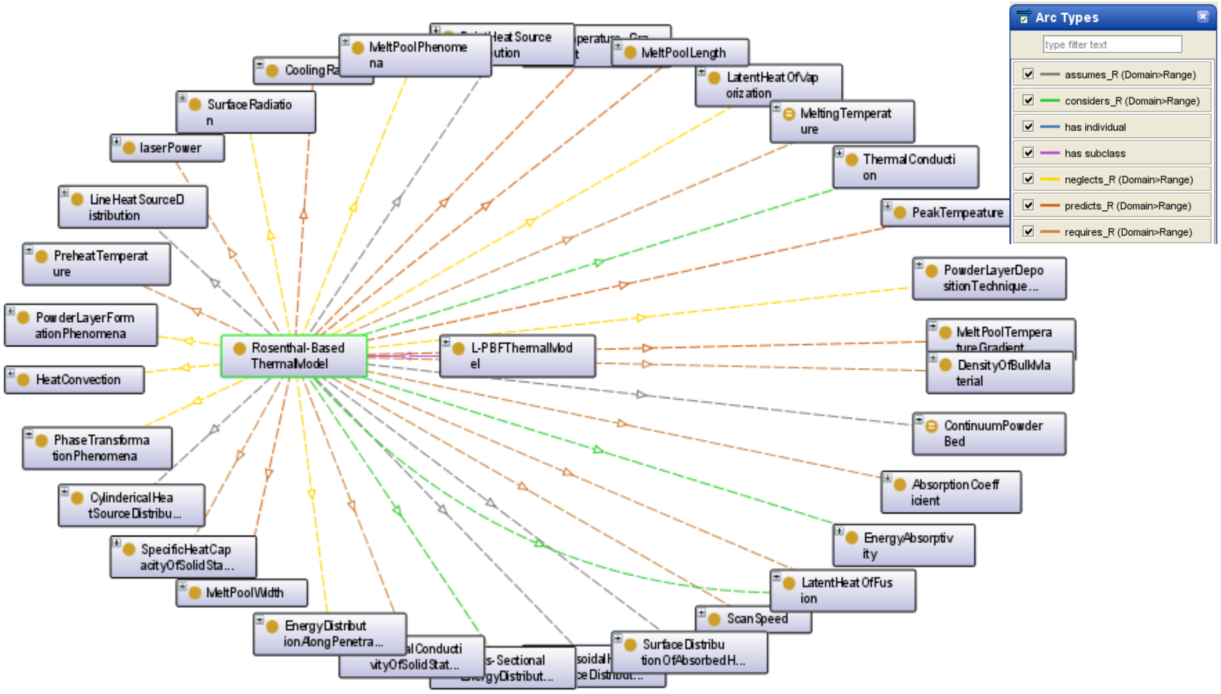
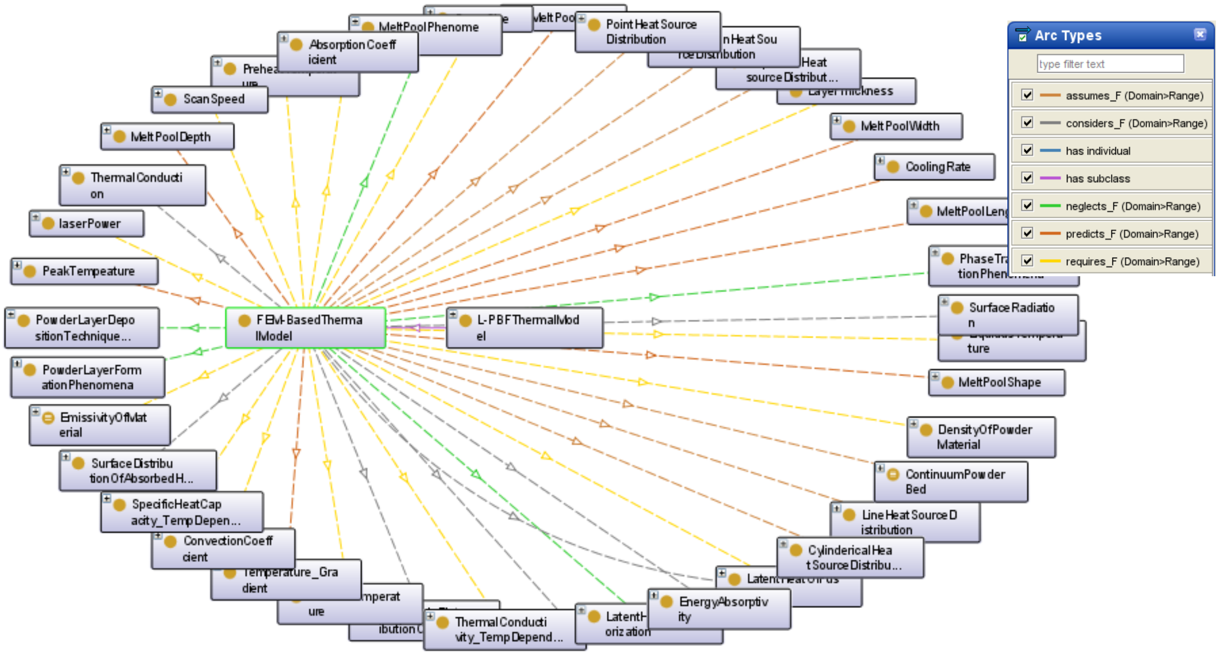

LPBFUncertainty
'LPBFUncertaintyQuantification(UQ)Method'
QuantifyingMeasurementUncertainty
QuantifyingModelFormUncertainty
QuantifyingNumericalUncertainty
QuantifyingParameterUncertainty
LPBFUncertaintySource
LPBFModelUncertaintySource
MiscellaneousUncertaintySources
'ModelFormUncertainty(MFU)'
MFUInHeatSourceModel
MFUInMeltPoolModel
MFUInPowderBedModel
MFUInResidualStressAndDistortionModel
MFUInSolidificationModel
NumericalUncertainty
DiscretizationError
FiniteNumberOfElements/Nodes
FiniteTimeStep
ErrorsDueToComputerProgrammingMistakes
IterationConvergenceError
Round-offError
SurrogateModelUncertainty
TruncationError
ParameterUncertainty
LaserPowerUncertaintySource
LayerThicknessUncertaintySource
OrientationError
BuildPlatformOrientationError
RecoaterArmOrientationError
PositionError
VariationInPowderBedDensity
Vibration
PowderBedAbsorptionCoefficientUncertaintySource
PowderBedEmissivityUncertaintySource
PowderBedThermalConductivityUncertaintySource
ScanSpeedUncertaintySource
MeasurementUncertainty
ErrorInInstrumentCalibration
Co-axialCameraCalibrationError
IRThermographyCalibrationError
OpticalCameraCalibrationError
PyrometryCalibrationError
ThermalCameraCalibrationError
ImpreciseMeasurementMethod
VariationInExperimentalMeasurements
TypeOfUncertanty
AleatoricUncertainty
CombinedUncertanty
EpistemicUncertainty

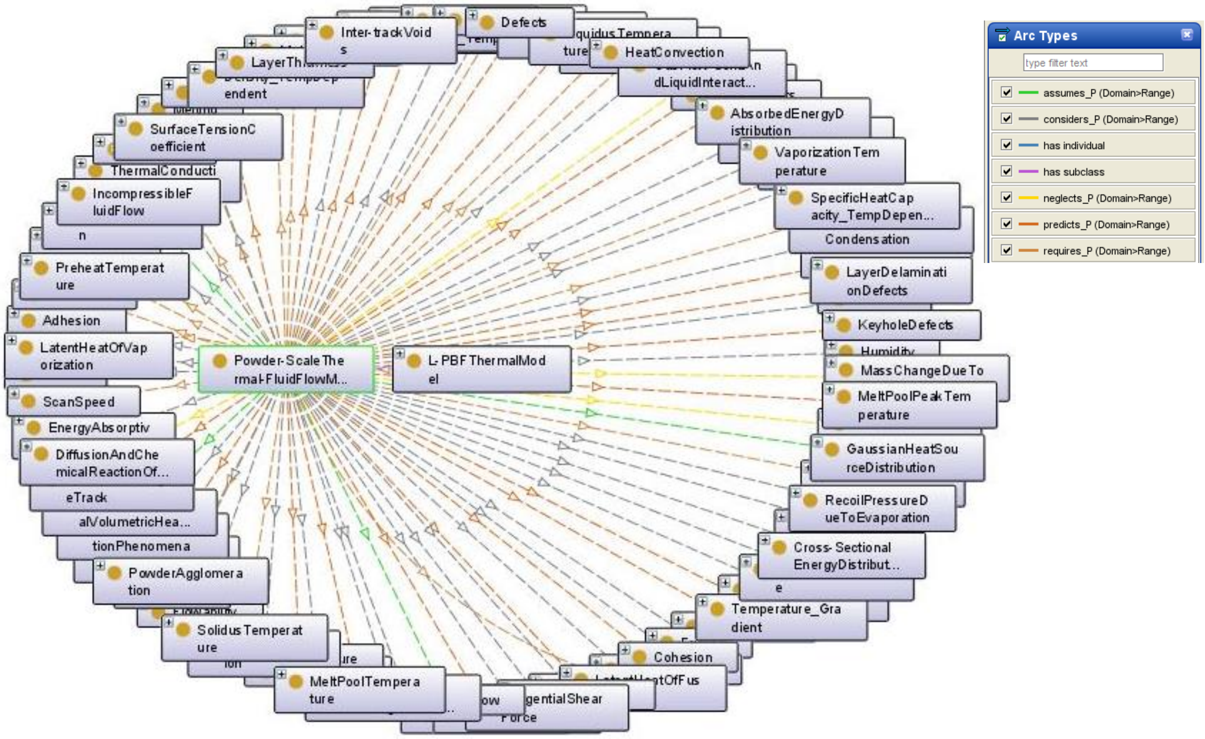
FIGURE 5: HIERARCHICAL REPRESENTATION OF L-PBF UNCERTAINTY SOURCES


Point Heat Source Distribution
...perature_Gra...t
MeltPoolPhenomena
Cooling Ra...
MeltPoolLength
LatentHeatOfVaporization
Surface Radiation
MeltingTemperature
laserPower
ThermalConduction
PeakTempeature
Line Heat Source Distribution
Preheat Temperature
PowderLayerDepositionTechnique...
PowderLayerFormationPhenomena
MeltPoolTemperatureGradient
Rosenthal-Based ThermalModel
L-PBFThermalModel
DensityOfBulkMaterial
HeatConvection
ContinuumPowderBed
PhaseTransformationPhenomena
AbsorptionCoefficient
CylindericalHeatSourceDistribu...
EnergyAbsorptivity
SpecificHeatCapacityOfSolidSta...
LatentHeatOfFusion
MeltPoolWidth
ScanSpeed
EnergyDistributionAlongPenetra...
...alConductivityOfSolidStat...
...s-Sectional EnergyDistribut...
...soidal H... ...e Distribut...
Surface Distribution Of Absorbed H...
Arc Types
type filter text
assumes_R (Domain>Range)
considers_R (Domain>Range)
has individual
has subclass
neglects_R (Domain>Range)
predicts_R (Domain>Range)
requires_R (Domain>Range)

(a) ONTOLOGICAL RELATIONSHIPS IN SEMI-ANALYTICAL ROSENTHAL-BASED THERMAL MODEL



(b) ONTOLOGICAL RELATIONSHIPS IN FEM-BASED THERMAL MODEL



(c) ONTOLOGICAL RELATIONSHIPS IN POWDER-SCALE THERMAL-FLUID FLOW MODEL

FIGURE 6: ONTOLOGICAL RELATIONSHIPS IN L-PBF AM MODELS

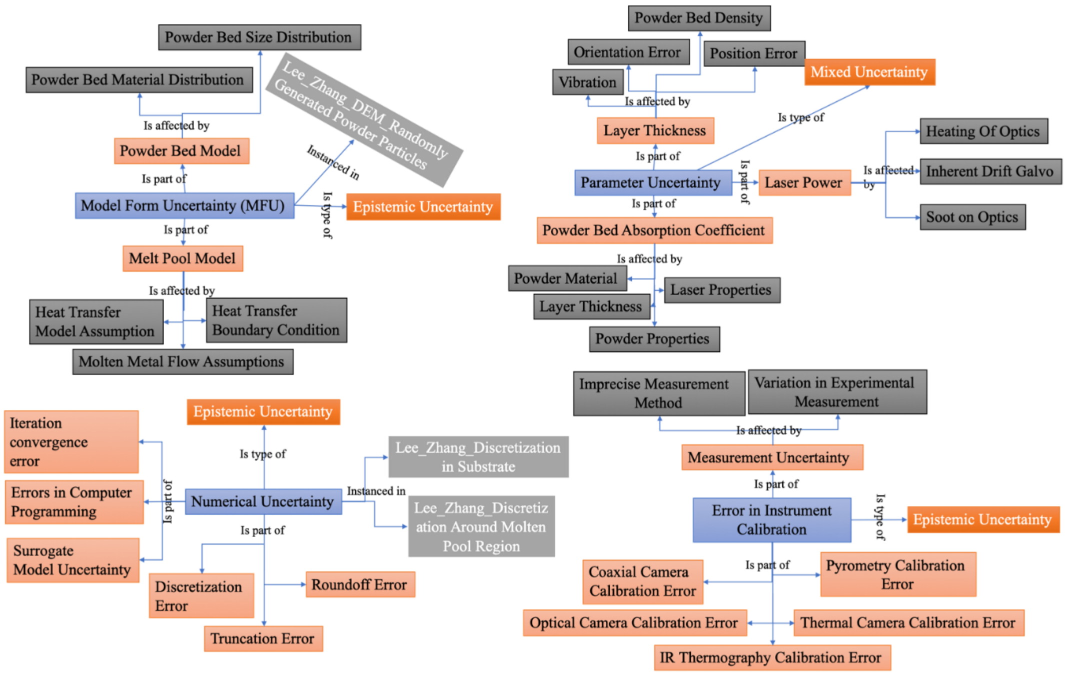

Powder Bed Size Distribution
Powder Bed Material Distribution
Lee_Zhang_DEM_Randomly Generated Powder Particles
Is affected by
Powder Bed Model
Instanced in
Is part of
Model Form Uncertainty (MFU)
Is type of
Epistemic Uncertainty
Is part of
Melt Pool Model
Is affected by
Heat Transfer Model Assumption
Heat Transfer Boundary Condition
Molten Metal Flow Assumptions
Powder Bed Density
Orientation Error
Position Error
Vibration
Mixed Uncertainty
Is affected by
Is type of
Layer Thickness
Heating Of Optics
Is part of
Is part of
Parameter Uncertainty
Laser Power
Is affected by
Inherent Drift Galvo
Is part of
Soot on Optics
Powder Bed Absorption Coefficient
Is affected by
Powder Material
Laser Properties
Layer Thickness
Powder Properties
Iteration convergence error
Epistemic Uncertainty
Is type of
Lee_Zhang_Discretization in Substrate
Errors in Computer Programming
Is part of
Numerical Uncertainty
Instanced in
Lee_Zhang_Discretiz ation Around Molten Pool Region
Surrogate Model Uncertainty
Is part of
Discretization Error
Roundoff Error
Truncation Error
Imprecise Measurement Method
Variation in Experimental Measurement
Is affected by
Measurement Uncertainty
Is part of
Error in Instrument Calibration
Is type of
Epistemic Uncertainty
Is part of
Coaxial Camera Calibration Error
Pyrometry Calibration Error
Optical Camera Calibration Error
Thermal Camera Calibration Error
IR Thermography Calibration Error

FIGURE 7: SOURCES OF MODEL FORM (TL), NUMERICAL (BL), PARAMETER (TR), AND MEASUREMENT (BR) UNCERTAINTIES

7. CONCLUSIONS

This study explored two major aspects of L-PBF additive manufacturing process. First, the L-PBF thermal models were characterized focusing on their input parameters, output QoIs, assumptions, and considered and neglected phenomena. These models include Rosenthal-based thermal models, FEM-based thermal models, and powder-scale thermal-fluid flow models. The characterization of the models is necessary to understand the abstraction and formulations of the models and investigate model elements that can be used as a basis for studying the sources of uncertainty which ultimately leads to understanding the predictive accuracy of the models. Then, characterization of sources of uncertainty of the L-PBF models was conducted. These uncertainty sources include model form, numerical, input parameters, and measurement uncertainties. Model form uncertainty is caused by assumptions and simplifications taken during representation of physical phenomena using governing mathematical equations. A validation approach can be used to characterize this source of uncertainty along with model bias by comparing simulation and measurement results. Discretization error associated with element size and time steps is the major cause of numerical uncertainty and a verification approach can be used to characterize and estimate the value of this uncertainty. The presence of natural variability in input variables and imprecise values cause uncertainty in output QoIs. Parameter uncertainty of a model can be determined by using sampling methods either directly on the model, provided that the model is computationally efficient, or through surrogate models that represent the computationally expensive physics-based models. Finally, measurement uncertainty, which is used in model validation, is caused by calibration error, imprecise measurement method, and variation in measurement results, and a standard approach is used to characterize this uncertainty.

This paper also presented ontological representation of the L-PBF models and uncertainty sources. The ontology based on Protégé captures the relevant knowledge associated to process, model, and uncertainty sources. This ontology represents the characterization of the powder-scale thermal-fluid flow models that better imitate the L-PBF process. The class of the L-PBF process mainly captures the physical characteristics, process parameters, and process signatures. Whereas the model ontology class captures model assumptions, formulations, input parameters, outputs, and predictive methods. The L-PBF uncertainty class captures sources of uncertainty, type of uncertainty, and uncertainty quantification methods and approaches. This work can be further extended by incorporating different models ranging from low to high-fidelity in order to capture more knowledge to assess their predictive capability and compare different models. The topology and mapping of the uncertainty sources presented in this study establish fundamental requirements for measuring model fidelity, and for guiding the selection of a model suitable for its intended purpose.

DISCLAIMER

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial equipment, instruments or materials are identified in this report to facilitate better understanding. Such identification does not imply recommendations or endorsement by NIST nor does it imply the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] Ian, G., David, R., and Brent, S., 2013, *Additive Manufacturing Technologies*, Springer.

[2] Petrovic, V., Vicente Haro Gonzalez, J., Jordá Ferrando, O., Delgado Gordillo, J., Ramón Blasco Puchades, J., and Portolés Griñan, L., 2011, “Additive Layered Manufacturing: Sectors of Industrial Application Shown through Case Studies,” Int. J. Prod. Res., **49**(4), pp. 1061–1079.

[3] Prashanth, K. G., Scudino, S., and Eckert, J., 2017, “Defining the Tensile Properties of Al-12Si Parts Produced by Selective Laser Melting,” Acta Mater., **126**, pp. 25–35.

[4] Bourell, D. L., Leu, M. C., and Rosen, D. W., 2009, “Roadmap for Additive Manufacturing: Identifying the Future of Freeform Processing,” Rapid Prototyp. J., **5**(4), pp. 169–178.

[5] Assouroko, Ibrahim; Lopez, Felipe; Witherell, P., 2016, “A Method for Characterizing Model Fidelity in Laser Powder Bed Fusion Additive Manufacturing,” *Proceedings of the ASME 2016 International Mechanical Engineering Congress & Exposition ASME IMECE 2016 November 11-17, 2016, Phoenix, Arizona, USA*, pp. 1–13.

[6] Smith, J., Xiong, W., Yan, W., Lin, S., Cheng, P., Kafka, O. L., Wagner, G. J., Cao, J., and Liu, W. K., 2016, “Linking Process, Structure, Property, and Performance for Metal-Based Additive Manufacturing: Computational Approaches with Experimental Support,” Comput. Mech., **57**(4), pp. 583–610.

[7] DebRoy, T., Wei, H. L., Zuback, J. S., Mukherjee, T., Elmer, J. W., Milewski, J. O., Beese, A. M., Wilson-Heid, A., De, A., and Zhang, W., 2018, “Additive Manufacturing of Metallic Components – Process, Structure and Properties,” Prog. Mater. Sci., **92**, pp. 112–224.

[8] Moges, T., Ameta, G., and Witherell, P., 2019, “A Review of Model Inaccuracy and Parameter Uncertainty in Laser Powder Bed Fusion Models and Simulations,” J. Manuf. Sci. Eng., **141**(4), p. 040801.

[9] Energetics Incorporated, 2013, *Measurement Science Roadmap for Metal-Based Additive Manufacturing*.

[10] Rosenthal, D., 1946, “The Theory of Moving Sources of Heat and Its Application to Metal Treatments,” Trans. ASME, pp. 849–866.

[11] King, W. E., Anderson, A. T., Ferencz, R. M., Hodge, N. E., Kamath, C., Khairallah, S. A., and Rubenchik, A. M., 2015, “Laser Powder Bed Fusion Additive Manufacturing of Metals; Physics, Computational, and Materials Challenges,” Appl. Phys. Rev., **2**(4), p. 041304.

[12] Gholaminezhad, I., Assimi, H., Jamali, A., and Vajari, D. A., 2016, “Uncertainty Quantification and Robust Modeling of Selective Laser Melting Process Using Stochastic Multi-Objective Approach,” Int. J. Adv. Manuf. Technol., **86**(5), pp. 1425–1441.

[13] Adamczak, S., Bochnia, J., and Kaczmarska, B., 2014, “Estimating the Uncertainty of Tensile Strength Measurement for a Photocured Material Produced by Additive Manufacturing,” Metrol. Meas. Syst., **21**(3), pp. 553–560.

[14] Hu, Z., and Mahadevan, S., 2017, “Uncertainty Quantification and Management in Additive Manufacturing: Current Status, Needs, and Opportunities,” Int. J. Adv. Manuf. Technol., **93**, pp. 2855–2874.

[15] Moges, T., Yan, W., Lin, S., Ameta, G., Fox, J., and Witherell, P.,

2018, “Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations,” Solid Free. Fabr. Symp., pp. 1913–1928.

[16] Lopez, F., Witherell, P., and Lane, B., 2016, “Identifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models,” J. Mech. Des., **138**(November), pp. 1–4.

[17] Moser, D., Beaman, J., Fish, S., and Murthy, J., 2014, “MULTI-LAYER COMPUTATIONAL MODELING OF SELECTIVE LASER SINTERING PROCESSES,” Proc. ASME 2014 Int. Mech. Eng. Congr. Expo. IMECE2014, pp. 1–11.

[18] Ma, L., Fong, J., Lane, B., Moylan, S., Filliben, J., Heckert, A., and Levine, L., 2015, “Using Design of Experiments in Finite Element Modeling To Identify Critical Variables for Laser Powder Bed Fusion,” Solid Free. Fabr. Symp., pp. 219–228.

[19] Nath, P., Hu, Z., and Mahadevan, S., 2017, “Multi-Level Uncertainty Quantification in Additive Manufacturing,” Solid Free. Fabr. 2017 Proc. 28th Annu. Int. S, pp. 922–937.

[20] Kamath, C., 2016, “Data Mining and Statistical Inference in Selective Laser Melting,” Int. J. Adv. Manuf. Technol., **86**(5–8), pp. 1659–1677.

[21] Eagar, T. W., and Tsai, N. S., 1983, “Temperature Fields Produced by Traveling Distributed Heat Sources,” Weld. J., **62**(12), pp. 346–355.

[22] Verhaeghe, F., Craeghs, T., Heulens, J., and Pandelaers, L., 2009, “A Pragmatic Model for Selective Laser Melting with Evaporation,” Acta Mater., **57**(20), pp. 6006–6012.

[23] Tapia, G., King, W. E., Arroyave, R., Johnson, L., Karaman, I., and Elwany, A., 2018, “Uncertainty Propagation Analysis of Computational Models in Laser Powder Bed Fusion Additive Manufacturing Using Polynomial Chaos Expansions,” J. Manuf. Sci. Eng., **140**(December).

[24] Tapia, G., Khairallah, S., Matthews, M., King, W. E., and Elwany, A., 2018, “Gaussian Process-Based Surrogate Modeling Framework for Process Planning in Laser Powder-Bed Fusion Additive Manufacturing of 316L Stainless Steel,” Int. J. Adv. Manuf. Technol., **94**(9–12), pp. 3591–3603.

[25] Khairallah, S. A., Anderson, A. T., Rubenchik, A., and King, W. E., 2016, “Laser Powder-Bed Fusion Additive Manufacturing: Physics of Complex Melt Flow and Formation Mechanisms of Pores, Spatter, and Denudation Zones,” Acta Mater., **108**, pp. 36–45.

[26] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., and Lopez, F., 2016, “Investigating Predictive Metamodeling for Additive Manufacturing,” Proc. ASME Des. Eng. Tech. Conf., **1A-2016**, pp. 1–10.

[27] Yang, Z., Yan, L., Yeung, H., and Krishnamurty, S., 2019, “Investigation of Deep Learning for Real-Time Melt Pool Classification in Additive Manufacturing,” *Proceedings of the 2019 IEEE International Conference on Automation Science and Engineering (CASE), Accepted.*, Vancouver, BC, Canada.

[28] Ko, H., Lu, Y., Witherell, P., and Ndiaye, N. Y., 2019, “Machine Learning Based Continuous Knowledge Engineering for Additive Manufacturing,” *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), Accepted.*, Vancouver, BC, Canada.

[29] Razvi, S. S., Feng, S., Narayanan, A., Lee, Y. T., and Witherell, P. “A, 2019, “A Review of Machine Learning Applications in Additive Manufacturing,” *Proceedings of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Accepted*, Anaheim, CA, USA.

[30] Hu, Z., and Mahadevan, S., 2017, “Uncertainty Quantification in Prediction of Material Properties during Additive Manufacturing,” Scr. Mater., **135**, pp. 135–140.

[31] Mahmoudi, M., Tapia, G., Karayagiz, K., Franco, B., Ma, J., Arroyave, R., Karaman, I., and Elwany, A., 2018, “Multivariate Calibration and Experimental Validation of a 3D Finite Element Thermal Model for Laser Powder Bed Fusion Metal Additive Manufacturing,” Integr. Mater. Manuf. Innov., **7**(3), pp. 116–135.

[32] Ghosh, S., Mahmoudi, M., Johnson, L., Elwany, A., Arroyave, R., and Allaire, D., 2019, “Uncertainty Analysis of Microsegregation during Laser Powder Bed Fusion,” Model. Simul. Mater. Sci. Eng., **27**(3), p. 034002.

[33] Zeng, K., Pal, D., and Stucker, B. E., 2012, “A Review of Thermal Analysis Methods in Laser Sintering and Selective Laser Melting,” Proc. Solid Free. Fabr. Symp., pp. 796–814.

[34] Schoinochoritis, B., Chantzis, D., and Salonitis, K., 2014, “Simulation of Metallic Powder Bed Additive Manufacturing Processes with the Finite Element Method: A Critical Review,” Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., **231**(1), pp. 96–117.

[35] Roberts, I. A., Wang, C. J., Esterlein, R., Stanford, M., and Mynors, D. J., 2009, “A Three-Dimensional Finite Element Analysis of the Temperature Field during Laser Melting of Metal Powders in Additive Layer Manufacturing,” Int. J. Mach. Tools Manuf., **49**(12–13), pp. 916–923.

[36] Yan, W., Ge, W., Qian, Y., Lin, S., Zhou, B., Liu, W. K., Lin, F., and Wagner, G. J., 2017, “Multi-Physics Modeling of Single/Multiple-Track Defect Mechanisms in Electron Beam Selective Melting,” Acta Mater., **134**, pp. 324–333.

[37] Lee, Y. S., and Zhang, W., 2016, “Modeling of Heat Transfer, Fluid Flow and Solidification Microstructure of Nickel-Base Superalloy Fabricated by Laser Powder Bed Fusion,” Addit. Manuf., **12**, pp. 178–188.

[38] Mukherjee, T., Wei, H. L., De, A., and DebRoy, T., 2018, “Heat and Fluid Flow in Additive Manufacturing—Part I: Modeling of Powder Bed Fusion,” Comput. Mater. Sci., **150**(February), pp. 304–313.

[39] Yan, W., Lin, S., Kafka, O. L., Lian, Y., Yu, C., Liu, Z., Yan, J., Wolff, S., Wu, H., Ndip-Agbor, E., Mozaffar, M., Ehmann, K., Cao, J., Wagner, G. J., and Liu, W. K., 2018, “Data-Driven Multi-Scale Multi-Physics Models to Derive Process–Structure–Property Relationships for Additive Manufacturing,” Comput. Mech., **61**(5), pp. 521–541.

[40] Oberkampf, W. L., and Roy, C. J., 2010, *Verification and Validation in Scientific Computing*, Cambridge University Press, Cambridge.

[41] ASME-V&V-20, 2009, *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*, American Society of Mechanical Engineers.

[42] Roy, C. J., and Oberkampf, W. L., 2011, “A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing,” Comput. Methods Appl. Mech. Eng., **200**(25–28), pp. 2131–2144.

[43] Roy, C. J., 2005, “Review of Code and Solution Verification Procedures for Computational Simulation,” J. Comput. Phys., **205**(1), pp. 131–156.

[44] PTC 60 / ASME V&V10, 2005, *Guide for Verification and Validation in Computational Solid Mechanics*, The American Society of Mechanical Engineers.

[45] Roache, P., 2001, “Code Verification by the Method of Manufactured Solutions,” J. Fluids Eng., **124**(1), pp. 4–10.

[46] Riley, M., and Grandhi, R., 2011, “A Method for the Quantification of Model-Form and Parametric Uncertainties in Physics-Based Simulations,” *52nd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*, American

Moges, Tesfaye; Witherell, Paul; Ameta, Gaurav. ”On Characterizing Uncertainty Sources in Laser Powder Bed Fusion Additive Manufacturing Models.” Paper presented at International Mechanical Engineering Congress and Exposition (IMECE2019), Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

Institute of Aeronautics and Astronautics, Reston, Virigina, pp. 1–10.

[47] Mahesh, M., Lane, B., Donmez, A., Feng, S., Moylan, S., and Fesperman, R., 2015, “Measurement Science Needs for Real-Time Control of Additive Manufacturing Powder Bed Fusion Processes,” NISTIR 8036, Natl. Inst. Stand. Technol., pp. 1–50.

[48] Jcgm, J. C. F. G. I. M., 2008, “Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement,” Int. Organ. Stand. Geneva ISBN, **50**(September), p. 134.

[49] Devesse, W., De Baere, D., and Guillaume, P., 2014, “The Isotherm Migration Method in Spherical Coordinates with a Moving Heat Source,” Int. J. Heat Mass Transf., **75**, pp. 726–735.

[50] Fox, J. C., Lane, B. M., and Yeung, H., 2017, “Measurement of Process Dynamics through Coaxially Aligned High Speed Near-Infrared Imaging in Laser Powder Bed Fusion Additive Manufacturing,” Proc. SPIE 10214, Thermosense Therm. Infrared Appl. XXXIX, **1**(301), p. 1021407.

[51] Kim, S., Rosen, D. W., Witherell, P., and Ko, H., 2018, “A Design for Additive Manufacturing Ontology to Support Manufacturability Analysis,” Vol. 2A 44th Des. Autom. Conf., p. V02AT03A036.

[52] Roh, B., Kumara, S. R. T., Simpson, T. W., and Witherell, P., 2016, “Ontology-Based Laser and Thermal Metamodels for Metal-Based Additive Manufacturing,” Proc. ASME 2016 Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf. IDETC/CIE 2016 August 21-24, 2016, Charlotte, North Carolina, pp. 1–8.

Moges, Tesfaye; Witherell, Paul; Ameta, Gaurav. ”On Characterizing Uncertainty Sources in Laser Powder Bed Fusion Additive Manufacturing Models.” Paper presented at International Mechanical Engineering Congress and Exposition (IMECE2019), Salt Lake City, UT, US. November 11, 2019 - November 14, 2019.

Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y. - J. Son, eds.

DIGITAL TWIN FOR SMART MANUFACTURING: THE SIMULATION ASPECT

Guodong Shao

Engineering Laboratory
National Institute of Standards and Technology
100 Bureau Drive
Gaithersburg, MD 20899, USA

Sanjay Jain

Department of Decision Sciences
The George Washington University
2201 G Street NW, Suite 415
Washington, DC- 20052, USA

Christoph Laroque

University of Applied Sciences Zwickau
Institute for Management and Information
Chair for Business Computing
Scheffelstrasse 39
08056 Zwickau, GERMANY

Loo Hay Lee

Department of Industrial and Systems Engineering
National University of Singapore
10 Kent Ridge Crescent
Singapore, 119260

Peter Lendermann

D-SIMLAB Technologies Pte. Ltd.
8 Jurong Town Hall Road
#23-05 JTC Summit
Singapore, 609434

Oliver Rose

Department of Computer Science
Universität der Bundeswehr München
Werner-Heisenberg-Weg 39
Neubiberg, 85577, GERMANY

ABSTRACT

The purpose of this panel is to discuss the state of the art in digital twin for manufacturing research and practice from the perspective of the simulation community. The panelists come from the US, Europe, and Asia representing academia, industry, and government. This paper begins with a short introduction to digital twins and then each panelist provides preliminary thoughts on concept, definitions, challenges, implementations, relevant standard activities, and future directions. Two panelists also report their digital twin projects and lessons learned. The panelists may have different viewpoints and may not totally agree with each other on some of the arguments, but the intention of the panel is not to unify researchers' thinking, but to list the research questions, initiate a deeper discussion, and try to help researchers in the simulation community with their future study topics on digital twins for manufacturing.

1 INTRODUCTION

Recent technology advancement of smart sensors, Internet of Things (IoT), cloud computing, Artificial Intelligence (AI), Cyber-Physical Systems (CPS), and modeling and simulation make it possible to realize the "digital twin" of a manufacturing product, system, and process (Bolton 2016). These technologies enable better real-time data collection, computation, communication, integration, modeling, simulation, optimization, and control that are required by digital twins. "Digital Twin" has become an important

Shao, Guodong; Jain, Sanjay; Laroque, Christoph; Lee, Loo Hay; Lendermann, Peter; Rose, Oliver. "Digital Twin for Smart Manufacturing: the Simulation Aspect." Paper presented at Winter Simulation Conference 2019, National Harbor, MD, US. December 08, 2019 - December 11, 2019.

component in programs and initiatives related to Smart Manufacturing, Digital Manufacturing, Advanced Manufacturing, and Industry 4.0 globally. It is a “hot” topic among researchers, educators, software vendors, and practitioners in these fields, as one panelist indicates that searches of the key word “digital twin” has been growing rapidly since 2016. On Gartner’s 2017 Hype Cycles of Emerging Technologies, digital twin is listed with a time to acceptance of (five to ten) years, i.e., one-half of companies, by 2022, will be using digital twins to achieve more efficient system performance analysis and improved productivity (Panetta 2017). The International Data Corporation (IDC) forecasts that companies investing in digital twins will see improvements of 30% in cycle times of their critical processes in the next five years.

However, manufacturers are not implementing or embracing digital twins as rapidly and efficiently as expected. This may be because digital twins are still in their infancy stage, and there is a lot of confusion about what they actually are, what they should include, and where to start to implement them. The lack of consensus among researchers and practitioners in different communities and different industrial sectors also hinders the acceptance of digital twins by manufacturers. Many companies, especially small- and medium-sized enterprises (SMEs), do not have the expertise and resources required to study and understand the digital twin concept, definitions, and associated challenges; and then effectively implement the digital twin concept for their products and manufacturing operations. They typically have neither sufficient information on the required technologies and standards, nor systematic procedures guiding the implementation of a digital twin. In the simulation community, we thought that we knew digital twins better because we have been performing modeling and simulation for a few decades. However, with the opportunities of new technologies and data and the challenges and requirements of new data-driven and real-time modeling, we, as a community, should equip and update ourselves for this new era of modeling and simulation.

The goal of this panel is to start a discussion regarding the state of the art in digital twins for manufacturing research and development from the perspective of the simulation community. The panelists come from the US (Guodong Shao and Sanjay Jain), Europe (Christoph Laroque and Oliver Rose), and Asia (Loo Hay Lee and Peter Lendermann). Among them are four researchers from academia (Sanjay Jain, Christoph Laroque, Oliver Rose, and Loo Hay Lee), one panelist from the US government (Guodong Shao), and one panelist from a software vendor (Peter Lendermann). Each panelist has provided preliminary thoughts on concept, definitions, challenges, implementations, and future directions. Two panelists also report on their digital twin projects and lessons learned. The panelists may have different viewpoints and may not totally agree with each other on some of the arguments, but the intention of the panel is not to unify researchers’ thinking, but to identify research questions, initiate a deeper discussion, and try to help researchers in the simulation community for their future study topics on digital twin for manufacturing.

The remainder of this paper contains the list of panelists’ statements, which represent their personal thoughts, their research findings, and their implementation results of digital twins.

2 PANELIST STATEMENTS

This section provides initial thoughts of each panelist on the simulation aspect of Digital Twin for Smart Manufacturing.

Guodong Shao

a. What is a digital twin?

The digital twin concept originated by Grieves in 2002 is to create a digital informational construct of a physical system as an entity on its own. This digital information would be a “twin” of the information that was embedded within the physical system and be linked with that physical system through the entire lifecycle of the system (Grieves and Vickers 2016). The digital twin concept allows manufacturers to create models of their production systems and processes using real-time data collected from smart sensors and used for near-real-time analysis and control. The digital twin and the physical system are connected through IoT or smart sensors and actuators. Synchronization between the digital twin and its physical twin, either online or offline, ensures that the production systems are constantly optimized as the digital twin receives real-time performance information from the physical system.

Currently, there are multiple different definitions of the digital twin out there (Ahuett-Garza and Kurfess 2018; Coronado et al. 2018; Garetti et al. 2012; GE 2018; Haag and Anderl 2018; Hughes 2018; Negri et al. 2017; Siemens 2018; Tao et al. 2017). Many of the definitions imply that a digital twin is an identical virtual duplication of a physical entity or an entire system. However, from my perspective, there may be multiple digital twins each representing different focus, aspect, or view of the system, i.e., each digital twin application should have its own focus. A digital twin is context-dependent and could be a partial representation of a physical system, it may consist of only relevant data and models that are specifically designed for their intended purpose (Boschert and Rosen 2018; Shao and Kibira 2018).

b. What are the relationships between digital twins and simulation models?

Many people may think that simulation models are digital twins. The fact is that a digital twin can be a simulation model, but a simulation model may not necessarily be a digital twin. Digital models used in simulations often have the same type of sensor information and controls of a digital twin, but the information may be generated and manipulated within the simulation. The simulation may replicate what could happen in the real world, but not necessarily what is currently happening (Wong 2018). Kritzinger et al. (2018) propose a classification of digital models into three subcategories based on their level of data integration between the physical and digital counterparts: (1) Digital model: a digital representation of an existing or planned physical object without any form of automated data exchange between the physical and digital objects. Most of the current offline simulation models are this kind of digital model; (2) Digital shadow: a digital model with an automated one-way data flow between the physical and digital objects, e.g., a simulation model using real-time sensor data as inputs (Yang et al. 2017); (3) Digital twin: a digital model with bi-directional data flow between the physical and digital objects, e.g., a simulation model that uses real-time sensor data as inputs and updates some of the parameters of a manufacturing process or equipment.

c. Typical digital twin applications for smart manufacturing

Digital twins can be used to ensure information continuity throughout the entire product/system lifecycle; perform real-time monitoring, predict system behavior, provide production control and optimization; view, analyze, and control the state of products or processes; enable preventive maintenance, and realize virtual commissioning. The applications of the digital twin concept help reduce resource downtime, improve product throughput and quality, reduce manufacturing costs, and ensure operation safety. Advanced digital twins may update products in the field and provide service to end-user customer (Hughes 2018).

d. What are the research directions to promote digital twin applications in the simulation community?

Digital twins are gaining more attention but are still in their early stage. There are a lot of challenges that need to be overcome before manufactures can effectively, economically, and correctly implement digital twin technologies. Manufacturers, especially SMEs, need help interpreting the concepts, relevant standards and technology implementations. In the simulation community, we need to help solve issues related to data management including data collection, data processing, and data analytics; real-time model synchronization that guarantees the digital twin reflects the current status of its physical twin; model generation that includes automatic data driven model creation and standard-based model generation; and model verification, validation, uncertainty quantification (VVUQ) (Shao and Kibira 2018; Lugaresi and Matta 2018).

e. Current relevant standardization efforts

Useful standards for digital twin implementation include guidelines for consistently performing credible digital twin modeling and specifications that define the information models and data formats to enable the interoperability of data and models within digital twins. NIST researchers currently participating in the development and testing of multiple such standards. Two of them are listed below:

- ISO 23247 - Digital Twin Manufacturing Framework: is intended to provide a generic manufacturing digital twin development framework that can be instantiated for case-specific digital twin implementation. The standard will have four parts: (1) Overview and general principles, (2) Reference architecture, (3) Digital representation of physical manufacturing elements, and (4) Information exchange. The completed framework standard will provide guidelines, methods, and approaches for the development and implementation of manufacturing digital twins. It will also help facilitate the composability of models and interoperability among modules, provide examples of data collection, modeling and simulation, communication, integration, and applications of relevant standards. The framework will also enable the generation and management of common data and model components that most digital twins need to have to facilitate the reuse of these components. For example, a simulation components library or model template may be useful for composing and reusing components for future models. This standard is currently work-in-progress.
- The American Society of Mechanical Engineers (ASME) Verification and Validation (V&V) standards committee is developing best practices, general guidance, and a common language for verification, validation, and uncertainty quantification for computational modeling and simulation in advanced manufacturing. The guidelines for incorporating VVUQ for data-driven models and throughout model lifecycle are especially applicable to digital twin development. It will allow better traceability, improved verification and validation capability, and better model credibility.

Sanjay Jain

From Virtual Factory to Digital Twin? The panel members' inputs present a range of overlapping perspectives on digital twins in the context of manufacturing. All the perspectives appear to agree on some major aspects. All of us consider digital twins to have simulation models as the key platform and include interfaces to the real system and to analytics applications as part of the concept. Some of us include a few analytics capabilities as part of the digital twin. Some of us explicitly identify the capability to vary level of details and supporting the lifecycle of the manufacturing system. With that overall agreement, the views appear to diverge a bit as we get into some details.

The challenge appears to be in achieving an alignment in our understanding at the deeper level. Considering that all the panel members are long time participants of Winter Simulation Conference (WSC), a practitioner or even a researcher from outside the community may expect us to be quite well aligned. The differences in our perspectives underline the need to work towards a common understanding. If we, being a part of the same community over a long period, differ on the details, it is not surprising that a whole range of diverse viewpoints are found in the larger community of manufacturing practitioners and researchers.

Interestingly the challenge of developing a common understanding of the digital twin concept is rather similar to the challenge with the virtual factory concept. Based on Google Scholar searches the earliest mention of virtual factory appears to be by Fisher (1986) as below:

> "Perhaps the most important benefit that can be derived from the development of an intelligent factory design agent is the ability to create an electronic model of the factory for subsequent use by other KSs (knowledge-based systems) and problem solvers. This virtual factory would benefit, for example, redesign of a factory when a change in product line occurs because only change related information would need to be collected due to the a priori existence of a factory model."

It can be seen that this original idea of virtual factory as an "electronic model" of the real factory that can be updated is quite similar to at least some definitions of digital twins. This is indeed why the challenge of definition of virtual factory has been referred to rather than any of other multitude of concepts that suffer from overuse with varying definitions. We submit that the digital twin in the context of manufacturing is almost the same concept as virtual factory, at least with the definition that we are using now and that is somewhat enhanced version of original idea described in Jain (1995).

Virtual factory was conceptualized as going beyond the simulation of only the material flow and immediately associated resources and activities. The three major enhancements proposed were in taking an integrated view of multiple relevant aspects of the factory, developing the virtual factory in parallel with the development of a real factory through its life cycle, and simulating and analyzing at different resolution levels. The concept was more recently enhanced in Jain and Shao (2014) to include open standard based interfaces with data sources and with analytics capabilities and is shown in Figure 1.

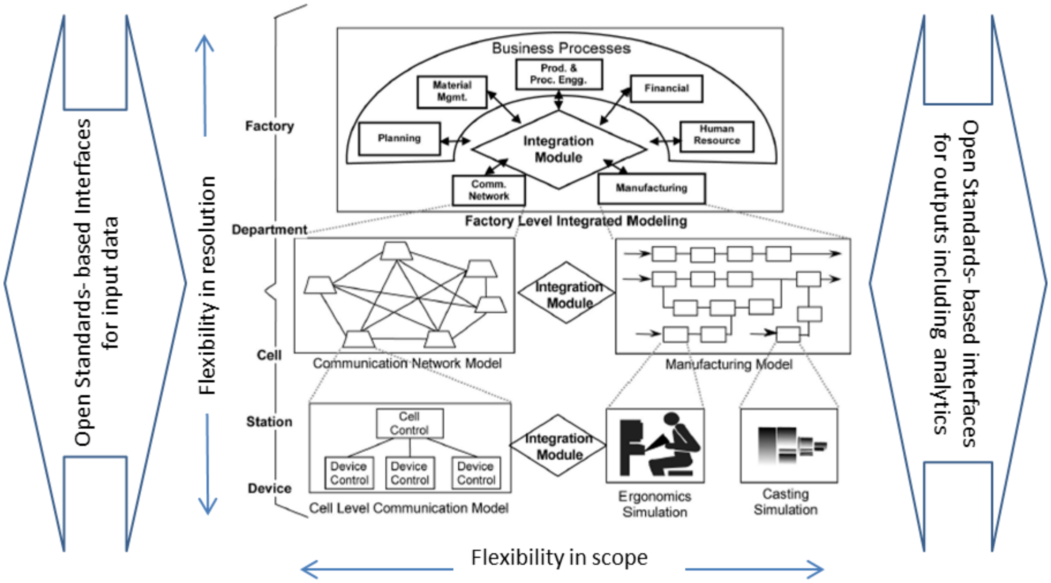


Figure 1: Virtual Factory concept (adapted from Jain and Shao 2014).

It should be apparent that the virtual factory concept largely overlaps with the digital twin concept applied to manufacturing. Digital twin is clearly a more generic concept as it can be applied to other environments such as a port and it appears to be used frequently for products. One would need to use an additional specifier such as the factory's digital twin. Some authors appear to use digital factory largely in the sense of factory's digital twins. It will be beneficial to all to agree on the terminology to avoid potential miscommunications between the providers and users of such capabilities.

It would help define not only one phrase representing the envisaged virtual factory or factory's digital twin capability, but also successively larger subsets that provide a path to start small and build a factory's true digital twin. The coining of digital model, shadow, and twin mentioned elsewhere in this paper is in the right direction and so is the idea of the increasing capabilities defined on four dimensions but perhaps a more comprehensive maturity model approach and/or additional dimensions are needed. Such a set would need to be developed via an international multi-party effort for wider acceptance. The development of the comprehensive model will help with better communication and allow practitioners and researchers to focus on advancing towards smart manufacturing without being lost in definitions.

The multi-resolution capability for the concept in Figure 1 will likely require multiple simulation paradigms for implementation including continuous simulation at for modeling individual manufacturing processes, discrete event simulation for modeling factory flow, and system dynamics for modeling interactions of business processes. Jain et al. (2015) present a virtual factory prototype that employs continuous simulation for modeling the turning process dynamics and kinematics, agent-based modeling for machine level model, and discrete event simulation for job shop level model. While the use of multiple paradigms provides the capability for analysis appropriate to level of detail, it does increase the expertise requirement for the modelers and analysts to carry out the task.

There are multiple challenges beyond definitions of the concept and the high expertise requirement for multiple resolution modeling that are facing manufacturers, particularly SMEs, interested in implementing their factory's digital twins. These include the effort and expertise required to collect and set up data for simulation, build the interfaces, analyze the outputs, and provide timely input to the decision makers. Technology advancements in multiple fields are helping address the challenges. Jain, Narayanan, and Lee (2019) propose a standards-based infrastructure to move towards addressing the challenges.

Christoph Laroque

The Digital Twin for Simulation in Operations – Something new beyond marketing? Data-driven Decision Support such as Simulation, Advanced Data Analytics, and AI are changing how modern manufacturing processes are planned and executed. Within the vision of Industry 4.0 and Cyber-Physical Production Systems, complex problems due to planning, scheduling and control of production, and logistic processes are derived by data-driven decisions in the nearer future. Thus, new processes and interoperable systems must be designed, and existing ones have to be improved, since Industry 4.0 has placed extremely high expectations on production systems to have substantial increase in productivity, resource efficiency, and level of automation. The deliverance of these expectations lies in the ability of manufacturing companies to accurately predict and plan their activities on the machine, the plant, as well as at the supply-chain-level.

Discrete event simulation (DES) is very suitable to model the reality in a manufacturing system with high fidelity. Such models are easy to parameterize and they are able to consider several influences including stochastic behavior. However, simulation models are challenged when it comes to operational decision support in manufacturing as well as logistics. The simulation models are very complex and need huge amount of production data and up to hours for the execution of simulation experiments. A better approach is to integrate the methods and algorithms from (big) data analytics and AI during the implementation of the "digital production twin" for different purposes, e.g., Predictive Maintenance or Workforce Scheduling. The digital twin represents the behavior of the corresponding real object or process and is compared with it at (mostly regular) defined points in time. A large amount of data can be used, the data is generated when implementing the Industry 4.0 concepts during operation as so-called "digital shadows."

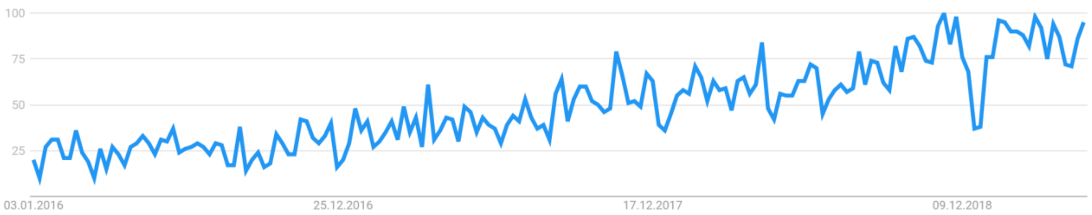


Figure 2: Worldwide Searches for the Term "Digital Twin" (Source: Google Trends).

One might say, that the concepts behind the innovative term “digital twin” might be old and known, which seems to be reasonably true from the perspective of a simulation expert. However, with the growing importance of searches for the term from all over the world (Figure2 indicates that searches grow by 400% in the last two years) and within the technological roadmap of the larger consultancies, the “digital twin”’ might lead to a higher visibility in top-management and at the decision-makers desk (at least this panelist thinks so).

But also, from a technological perspective, it might be reasonable to think about innovative combinations of the existing data-driven methods for decision making or decision support with DES, specifically material flow simulation, in order to implement more applications of simulation in daily manufacturing operations to achieve better planning, scheduling, and control results. Especially, approaches from data analytics that perform pre-simulation data aggregation, selection, and analysis might lead to performing successful applications in the manufacturing practice.

Loo Hay Lee

Building Toward the Digital Twin for the Smart System: A digital twin is the manifestation of the physical system in the digital world that can be used for various purposes. It can provide an environment for monitoring, testing, planning, and decision-making without real physical or time constraints. Besides the spatial representation of its physical counterpart, digital twin also needs to include simulation model and analytic methodologies.

The desired capability for the digital twin includes four dimensions as illustrated in Figure 3. Namely, the **Connectivity** that indicates the level of communication with its physical counterpart; the **Visibility** that indicates the ease of perception for human beings; the **Granularity** that indicates the detail level of the model, which can help us to look into the future scenarios in different fidelities; and the **Analyzability** that indicates how it can be used to assist for decision making (e.g., simulation optimization that can help us to find the best decision for the future; an analytics tool that can help us to learn based on the future simulated optimized data).

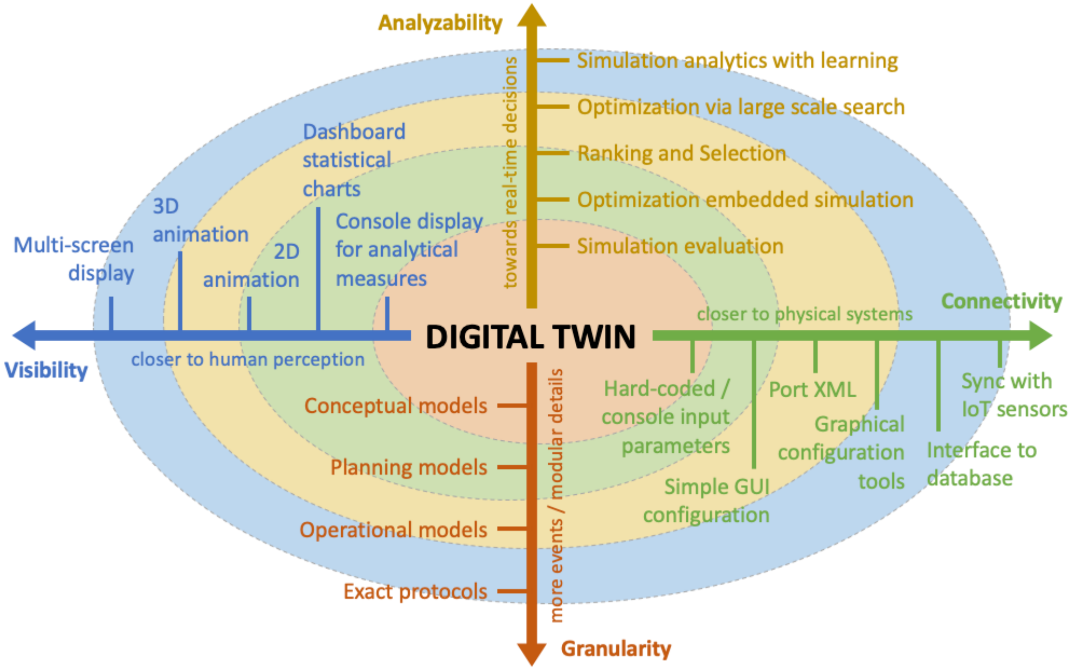

Analyzability
towards real-time decisions
Simulation analytics with learning
Optimization via large scale search
Ranking and Selection
Optimization embedded simulation
Simulation evaluation
Dashboard statistical charts
3D animation
Multi-screen display
2D animation
Console display for analytical measures
Visibility
closer to human perception
DIGITAL TWIN
closer to physical systems
Connectivity
Hard-coded / console input parameters
Port XML
Graphical configuration tools
Interface to database
Sync with IoT sensors
Simple GUI configuration
Conceptual models
Planning models
Operational models
Exact protocols
more events / modular details
Granularity

Figure 3: The four dimensions of desired capability for digital twin.

We have developed an O^{2}DES (object-oriented discrete-event simulation) framework as shown in Figure 4 (Zhou et al. 2017). With a rigorously defined Trinary modeling paradigm, the O^{2}DES framework allows developers and researchers to implement algorithmic tools to perform various types of analysis including (1) simulation to handle discrete event model, (2) optimization in simulation that can help to model the operation decision, (3) simulation in optimization (SimOpt approach) that can help to find the best decision under each scenario (Xu et al. 2015; Xu et al. 2016), as well as (4) learning based decision-making, i.e., simulation analytics that can learn the optimal decision function based on future optimized data. We have used this framework to develop digital twin for container terminal (Li et al. 2017; Zhou et al. 2018), aircraft spare part management (Li et al. 2015), warehouse (Pedrielli et al. 2016), and wafer fab plant.

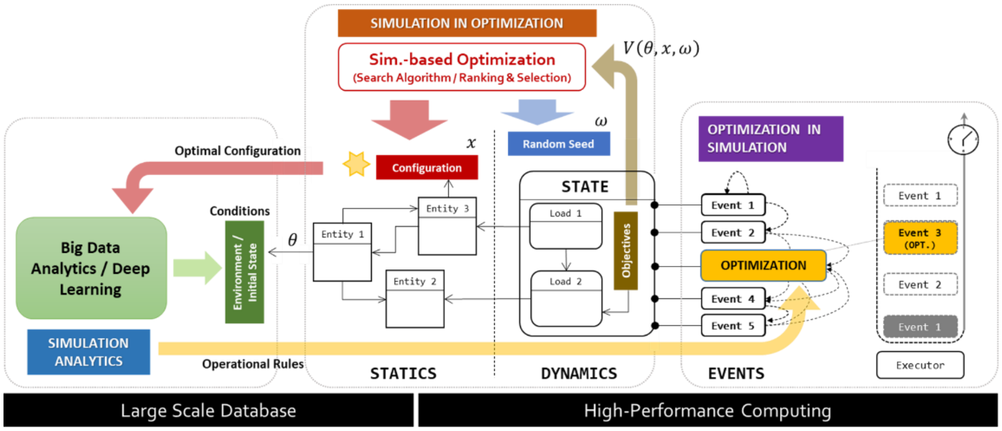


Figure 4: The illustration of O^{2}DES framework with trinary modeling paradigm.

Digital twins are not only the crystal ball to look into future but also the doctors that help provide solution for the future. Digital twins can enable us to actively learn from future, so that we are more prepared for the future, and aim to learn for success.

Peter Lendermann

Challenges with regard to the Usefulness of Digital Twins: The potential of the digital twin concept for the enhancement and continuous re-optimization of manufacturing and logistics operations has generally been recognized and accepted not only by academia but also by industry as it is an important backbone of the Industry 4.0 paradigm.

As mentioned by several co-panelists, simulation is an important enabler for creating a digital twin of a manufacturing and/or logistics system. However, a digital twin will never be able to be an “identical virtual duplication of a physical entity or an entire system” as stated by Shao, main reason being that the behavior of basically all manufacturing and logistics systems also involves human considerations and decision-making which inherently cannot be portrayed a computer simulation model. As such, the digital twin concept appears to be applicable mainly for highly automated systems with little human involvement.

In D-SIMLAB Technologies, the concept of digital twin is currently pursued mainly for semiconductor manufacturing, in particular highly automated 300 mm wafer fabs.

An additional complication in such a manufacturing environment, however, is the high degree of randomness on the production floor, caused by process steps such as quality measurement that, dependent on their outcome, may or may not result in re-work. As such, meaningful deterministic forecasts are only possible for very short time horizons in the order of a few hours at maximum.

Such deterministic forecasts are also the basis for complicated scheduling procedures that nowadays are used to optimize the material flow performance at critical equipment groups in wafer fabs. How this typically looks like in a wafer fab in terms of system architecture is outlined in the upper half of Figure 5.

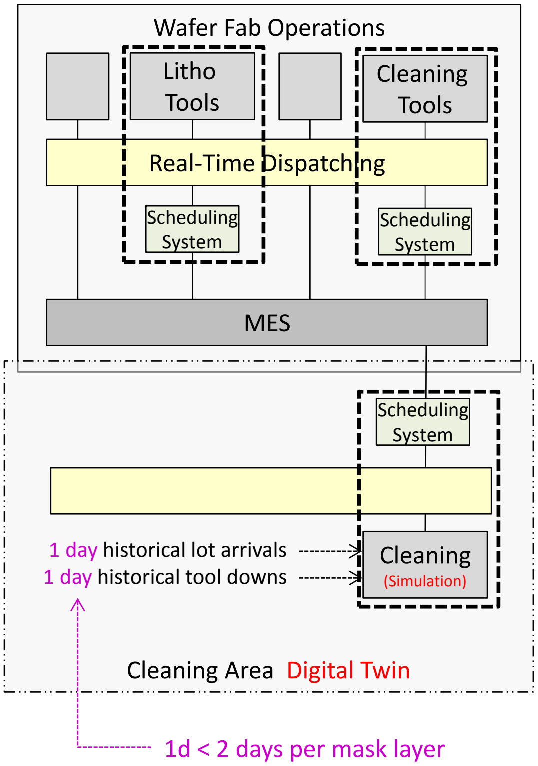


Figure 5: Simplified system architecture for material flow management in a wafer fab (upper half) and Digital Twin representing the cleaning area (lower half).

An important question to be addressed through a digital twin could be, for example whether certain scheduling parameters can be enhanced and better parameter values can be identified consistently. However, in a cleaning area of a large 300 mm fab comprising more than 100 wet benches, furnaces, and metrology tools, for example, commercially available scheduling tools typically run at a frequency of once every 10 min, whereby the scheduling procedure runs most of this time and the remaining time is needed for data input and output. This basically means that the scheduler runs almost continuously and hence also the digital twin, i.e., the simulation model of the cleaning area (in which the scheduler would have to run equally frequently) will inherently not be able to run faster than real-time. Optimization of scheduling parameters, in the sense of what are the best parameter values under which circumstances, will therefore be possible only retrospectively by comparing the (simulated) performance associated with different scheduler settings for different historical down or lot arrival patterns.

As indicated in Figure 6, parallel execution of different scenarios will be required, otherwise a meaningful analysis of scheduling parameters will not be possible. Also, multiple instances of the Scheduling solution will be required, basically equivalent to the number of instances that would be required to compare different scenarios on a cloud infrastructure, posing challenges to the licensing models currently practiced by commercial vendors of scheduling solutions.

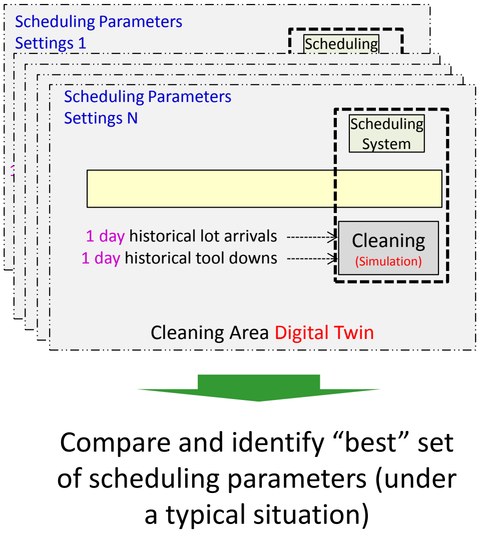

Scheduling Parameters
Settings 1
Scheduling
Scheduling Parameters
Settings N
Scheduling System
1 day historical lot arrivals
1 day historical tool downs
Cleaning
(Simulation)
Cleaning Area Digital Twin
Compare and identify "best" set of scheduling parameters (under a typical situation)

Figure 6: Comparison of different Digital Twin settings on a parallel (Cloud) computing infrastructure.

Oliver Rose

The term "Digital Twin" was coined as a part of the huge marketing campaign called Industrie/Industry 4.0 to speed up digitalization in production and logistics. Computer simulation models of production and logistics models are in use since the 50s of the last century. A computer model is a digital twin per se. What is more important is the question of what will be achieved with a newly invented digital twin that could not be achieved with computer simulation of manufacturing systems before. In my opinion, the goals are the same, the concepts and methods are the same, and eventually the problems are the same. For high-fidelity trustworthy models we needed and need appropriate data sources that still do not exist in almost all industries, even in high-tech cutting-edge manufacturing facilities, after decades of trying to achieve digital factories, smart factories, and the like. The only difference compared to the approaches of the past is that our computer equipment that is used to analyze the data, build models, and run simulations became much more capable over the years: it is considerably faster and has more memory. This means that we can have more model details, more simulation runs, and improved methods for analyzing data and building models such as machine learning. But this is just an evolution of the same old concept and nothing that is fundamentally new.

SUMMARY

In this panel paper, the panelists' statements are meant to aid researchers and manufacturers to have a better understanding of the concept, definition, challenges, and modeling requirements of digital twins. Two panelists also provided implementation examples to explain the digital twin concept and reported lessons learned. The panelists' statements represent their preliminary thoughts. The panelists may have different viewpoints, but all these viewpoints are worthwhile for further investigation and research.

This panel initiated a discussion on the topic of digital twin for smart manufacturing in the simulation community. The implementations of the digital twin concept have been initiated and better received in the design community for monitoring and improving product design (e.g., jet engine or turbines) and performance throughout the product lifecycle (Grieves 2014). It is mainly because of the characteristics of the problems and the relevant technological advances, i.e., the existence of the formal models of the products (e.g., CAD models) and the capabilities of integrating the system representation models with the system analysis models. These factors facilitate the successful implementation of digital twins for products.

The implementation of the digital twin concept in manufacturing has seen multiple approaches used with varying success as clear from the preceding statements of the panel members. The manufacturing community does not have the benefit of widely accepted formal models of the factory configuration or factory control, though there have been some efforts in this direction. The Core Manufacturing Simulation Data (CMSD) standard (Riddick and Lee 2010) developed under the auspices of Simulation Interoperability Standards Organization (SISO) has been used by multiple researchers in the US and Europe for representing factory configuration data with associated simulation data, such as statistical distributions for processing times. Lin and McGinnis (2017) show the feasibility of developing standard reference models for semantics and syntax of semiconductor manufacturing system models. Continued progress of such efforts will facilitate a more common approach for digital twins in manufacturing.

In the WSC community, most of the manufacturing applications are DES models of the manufacturing systems, processes, and supply chains. Some work does use multiple paradigms as pointed out by one of the panelists. Digital twins of manufacturing systems are anticipated to continue to largely use DES models. The simulation community is invited to help the move towards digital twins of manufacturing systems with efforts in the following tentative areas that are likely to get updated during the panel discussion at the conference:

- Agree on a definition of digital twins in manufacturing perhaps with an associated maturity model that allows clear identification of their capabilities at each stage.
- Agree on standard representations for configurations at each level of manufacturing hierarchy (e.g., machine, cell, line, factory, supply chain). The standards may vary for different manufacturing sectors.
- Develop and agree on standard representation of manufacturing control systems at different hierarchical levels. Again, the standards may vary for different manufacturing sectors.
- Enhance the capabilities of real-time model generation and its validation.
- Develop interfaces for model synchronization with its real manufacturing system counterpart.
- Develop standards for interfaces between models with different simulation paradigms.
- Integrate the digital twins with data analytics applications.
- Integrate digital twins' visualization capabilities (e.g., virtual reality(VR), augmented reality (AR), or mixed reality (MR)) if needed.
- Develop standard infrastructure for digital twins for manufacturing in particular for their implementations by small and medium enterprises.

DISCLAIMER

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

REFERENCES

Ahuett- Garza, H. and Kurfess, T. 2018. A brief discussion on the trends of habilitating technologies for industry 4.0 and smart manufacturing. Manufacturing Letters 15 (2018) 60–63.

Bolton, D. 2016. What Are Digital Twins and Why Will They Be Integral to The Internet Of Things? https://www.applause.com/blog/digital-twins-iot-faq/, accessed 15th April, 2019.

Boschert S, Rosen R. 2016. Digital Twin- The Simulation Aspect. In: Hehenberger P, Bradley D, editors. Mechatronic Futures. Cham: Springer International Publishing.

Coronado, P. D. U., Lynn, R., Louhichi, W., Parto, M., Wescoat, E., and Kurfess, T. 2018. Part data integration in the shop floor digital twin: mobile and cloud technologies to enable a manufacturing execution system. Journal of Manufacturing Systems 48 (2018) 25–33.

Fisher, E.L., 1986. "An AI-based Methodology for Factory Design." AI Magazine, 7(4), pp.72-72.

Garetti, M. Rosa, P. and Terzi, S. 2012. Life cycle simulation for the design of product-service systems. Computers in Industry, 63(2012) pp361-369. doi: 10.1016/j.compind.2012.02.007.

GE. 2018. What is Digital Twin? https://www.ge.com/digital/applications/digital-twin. accessed 15th April, 2019.

Grieves, M. 2014. "Digital Twin: Manufacturing Excellence through Virtual Replication". White Paper, http://innovate.fit.edu/plm/documents/doc_mgr/912/1411.0_Digital_Twin_White_Paper_Dr_Grieves.pdf, accessed 15th April, 2019.

Grieves, M. and J. Vickers. 2016. "Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems (Excerpt)", DOI:10.13140/RG.2.2.26367.61609, https://www.researchgate.net/publication/307509727_Origins_of_the_Digital_Twin_Concept. accessed 15th April, 2019.

Haag, S. and Anderl, R. 2018. Digital Twin – Proof of concept. Manufacturing Letters 15 (2018) 64–66.

Hughes, A. 2018. Forging the Digital Twin in Discrete Manufacturing, A Vision for Unity in the Virtual and Real Worlds. LNS Research. https://ifwe.3ds.com/sites/default/files/Forging%20the%20Digital%20Twin%20in%20Discrete%20Manufacturing.pdf. Accessed 15th April.

Jain, S. 1995. "Virtual Factory Framework: A Key Enabler for Agile Manufacturing." In *Proceedings 1995 INRIA/IEEE Symposium on Emerging Technologies and Factory Automation*. ETFA'95 (Vol. 1, pp. 247-258). IEEE.

Jain, S., D. Lechevalier, J. Woo, J. and S.J. Shin. 2015, "Towards a Virtual Factory Prototype." In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 2207-2218. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Jain, S., A. Narayanan, and Y.T. Lee. 2019. "Infrastructure for Model Based Analytics for Manufacturing." In *Proceedings of the 2019 Winter Simulation Conference,* edited by N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y-J. Son. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Jain, S. and G. Shao. 2014. "Virtual Factory Revisited for Manufacturing Data Analytics." In Proceedings of the 2014 Winter Simulation Conference, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 887-898. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., and Sihn, W. 2018. "Digital Twin in Manufacturing: A Categorical Literature Review and Classification". In *Paper Archive of the IFAC Conference Papers OnLine*. 51-11 (2018) 1016-1022. https://d1keuthy5s86c8.cloudfront.net/static/ems/upload/files/tcrze_0370_FI.pdf, accessed 15th April, 2019.

Li, H., Zhu, Y., Chen, Y., Pedrielli, G., and Pujowidianto, N. A. 2015. "The object-oriented discrete event simulation modeling: a case study on aircraft spare part management". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 3514-3525. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Li, H., Zhou, C., Lee, B. K., Lee, L. H., Chew, E. P., and Goh, R. S. M. 2017. "Capacity planning for mega container terminals with multi-objective and multi-fidelity simulation optimization". IISE Transactions, 49(9), 849-862.

Lugaresi, G. and Matta, A. 2018. "Real-Time Simulation in Manufacturing Systems: Challenges and Research Directions". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1226–1237. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Negri, E., Fumagalli, L., Macchi, M. 2017. A review of the roles of Digital Twin in CPS-based production systems. Procedia Manufacturing 11(2017) 939-948.Yang, W., Tan, Y., Yoshida, K., and Takakuwa, S. 2017. *DAAAM International Scientific Book 2017.* Chapter 18. Pp. 227-234.

Panetta, K. 2017. Top Trends in the Gartner Hype Cycle for Emerging Technologies. https://www.gartner. com/smarterwithgartner/top-trends-in-the-gartner-hype-cycle-for-emerging-technologies-2017/, accessed 15th April, 2019.

Pedrielli, G., Vinsensius, A., Chew, E. P., Lee, L. H., Duri, A., and Li, H. 2016. "Hybrid order picking strategies for fashion E-commerce warehouse systems". In *Proceedings of the 2016 Winter simulation conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 2250-2261. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Riddick, F.H., and Y.T. Lee. 2010. "Core Manufacturing Simulation Data (CMSD): A Standard Representation for Manufacturing Simulation-Related Information." In *Fall Simulation Interoperability Workshop (Fall SIW)* 20-24. SISO.

Shao, G. and Kibira, D. 2018. "Digital Manufacturing: Requirements and Challenges for Implementing Digital Surrogates". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1226–1237. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Siemens. 2018. Digital Twin. https://www.plm.automation.siemens.com/global/en/our-story/glossary/digital-twin/24465. accessed 15th April, 2019.

Tao, F. and Zhang, M. 2017. Digital Twin Shop-Floor: A New Shop-Floor Paradigm Towards Smart Manufacturing. IEEE Access, Special Section on Key Technologies for Smart Factory of Industry 4.0. Doi 10.1109/ACCESS.2017.2756069.

Wong, W. 2018. "What's the Difference Between a Simulation and a Digital Twin?" ElectricDesign. https://www.electronicdesign.com/embedded-revolution/what-s-difference-between-simulation-and-digital-twin. 15th April, 2019.

Xu, J., Huang, E., Chen, C. H., and Lee, L. H. 2015. "Simulation optimization: A review and exploration in the new era of cloud computing and big data". *Asia-Pacific Journal of Operational Research*, *32*(03), 1550019.

Xu, J., Huang, E., Hsieh, L., Lee, L. H., Jia, Q. S., and Chen, C. H. 2016. "Simulation optimization in the era of Industrial 4.0 and the Industrial Internet". *Journal of Simulation*, *10*(4), 310-320.

Zhou, C., Lee, L. H., Chew, E. P., and Li, H. 2017. "A modularized simulation for traffic network in container terminals via network of servers with dynamic rates". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 3150-3161. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhou, C., Li, H., Lee, B. K., and Qiu, Z. 2018. "A simulation-based vessel-truck coordination strategy for lighterage terminals". Transportation Research Part C: Emerging Technologies, 95, 149-164.

AUTHOR BIOGRAPHIES

GUODONG SHAO is a Computer Scientist in the Life Cycle Engineering Group in the Systems Integration Division (SID) of the Engineering Laboratory (EL) at the National Institute of Standards and Technology (NIST). His current research topics include digital twin; modeling, simulation, and analysis; data analytics; optimization; and model verification and validation for Smart Manufacturing. He holds a PhD in Information Technology. His email address is gshao@nist.gov.

SANJAY JAIN is an Associate Industry Professor in the Department of Decision Sciences, School of Business at the George Washington University. Before moving to academia, he accumulated over a dozen years of industrial R&D and consulting experience working at Accenture in Reston, VA, USA, Singapore Institute of Manufacturing Technology, Singapore and General Motors North American Operations Technical Center in Warren, MI, USA. His research interests are in application of modeling and simulation of complex scenarios including smart manufacturing systems and project management. His email address is jain@email.gwu.edu.

CHRISTOPH LAROQUE studied business computing at the University of Paderborn, Germany. Since 2013 he is Professor of Business Computing at the University of Applied Sciences Zwickau, Germany. He is mainly interested in the application of simulation-based decision support techniques for operational production and project management. His email address is Christoph.laroque@fh-zwickau.de.

LOO HAY LEE is an Associate Professor in the Department of Industrial and Systems Engineering, National University of Singapore. He received his B.S. (Electrical Engineering) degree from the National Taiwan University in 1992 and his S.M. and Ph.D. degrees in 1994 and 1997 from Harvard University. He is currently a senior member of IEEE, committee member of ORSS, and a member of INFORMS. His research interests include simulation-based optimization, production scheduling and sequencing and logistics and supply chain modeling. His email address is <iseleelh@nus.edu.sg> and his website is <www.ise.nus.edu.sg/staff/leelh/>.

PETER LENDERMANN is the Co-Founder and CEO of D-SIMLAB Technologies, a Singapore-based company providing simulation-based decision support solutions and services to Aviation and Semiconductor Manufacturing domains. Prior to this he worked at the Singapore Institute of Manufacturing Technology where he led the simulation-related research activities until spinning them off into D-SIMLAB Technologies. He has been engaged in the simulation community since the early 1990's. Peter holds a PhD in Applied High-Energy Physics from Humboldt University in Berlin (Germany) and an MBA in International Economics and Management from SDA Bocconi in Milan (Italy). His email address is peter@d-simlab.com.

OLIVER ROSE holds the Chair for Modeling and Simulation at the Department of Computer Science of the Universität der Bundeswehr, Germany. He received an M.S. degree in applied mathematics and a Ph.D. degree in computer science from Würzburg University, Germany. His research focuses on the operational modeling, analysis and material flow control of complex manufacturing facilities. He is a member of WSC Board of Directors, ASIM Board, and GI. His email address is oliver.rose@unibw.de.

Active Learning Yields Better Training Data for Scientific Named Entity Recognition

Roselyne B. Tchoua*, Aswathy Ajith*, Zhi Hong*, Logan T. Ward*‡,
Kyle Chard†‡, Debra J. Audus§, Shrayesh N. Patel¶, Juan J. de Pablo¶ and Ian T. Foster*†‡
*Department of Computer Science, University of Chicago, Chicago, IL, USA
†Globus, University of Chicago, Chicago, IL, USA
‡Data Science and Learning Division, Argonne National Laboratory, Argonne, IL, USA
§Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
¶Institute for Molecular Engineering, University of Chicago, Chicago, IL, USA
Email: roselyne@uchicago.edu

***Abstract*—Despite significant progress in natural language processing, machine learning models require substantial expert-annotated training data to perform well in tasks such as named entity recognition (NER) and entity relations extraction. Furthermore, NER is often more complicated when working with scientific text. For example, in polymer science, chemical structure may be encoded using nonstandard naming conventions, the same concept can be expressed using many different terms (synonymy), and authors may refer to polymers with ad-hoc labels. These challenges, which are not unique to polymer science, make it difficult to generate training data, as specialized skills are needed to label text correctly. We have previously designed polyNER, a semi-automated system for efficient identification of scientific entities in text. PolyNER applies word embedding models to generate entity-rich corpora for productive expert labeling, and then uses the resulting labeled data to bootstrap a context-based classifier. PolyNER facilitates a labeling process that is otherwise tedious and expensive. Here, we use active learning to efficiently obtain more annotations from experts and improve performance. Our approach requires just five hours of expert time to achieve discrimination capacity comparable to that of a state-of-the-art chemical NER toolkit.**

***Index Terms*—Named Entity Recognition, Machine Learning, Word Embedding, Active Learning, Polymers**

I. Introduction

A wealth of valuable research data is published in unstructured form in millions of scientific articles each year. Reading and extracting pertinent information from those articles has become an unmanageable task for scientists and makes it hard to build on existing results. A major obstacle to scientific fact extraction is the difficulty of identifying scientific entities in text. Despite much progress in natural language processing (NLP), scientific named entity recognition (NER) remains a research challenge. The main reason for this gap between NLP advances and scientific extraction needs is the lack of carefully annotated datasets for specific targets. In standard NER, progress is made possible by, for example, the Conference on Computational Natural Language Learning (CoNLL) dataset, which supports much work that advances the state of the art. But NER systems trained on CoNLL data do not perform well for scientific text, due to the distinctive vocabularies used in different scientific disciplines and subdisciplines.

Science-specific training datasets have been established in biology [1] and, more recently, chemistry [2]. However, the expert effort required to design extraction schema, define clear annotation rules, and generate training data is substantial, and cannot feasibly be performed for every field of science. As a consequence, annotated datasets do not exist in most fields, preventing the application of state-of-the-art NER and fact extraction methods.

This problem is apparent in materials science, where materials informatics seeks to combine large datasets and computational models to replace current trial-and-error materials design processes with targeted materials design, thus reducing time-to-market and development costs of new materials [3]. Unfortunately, the lack of annotated training data has hindered progress, preventing large-scale application of NER methods pioneered in, for example, biomedicine. Those methods, which often use hybrid rule-based, machine learning, and statistical techniques to extract entity names and relations from the literature [4, 5], require much training data. While similar efforts have begun in materials science [6–10], the lack of available training data impedes rapid progress. Instead, each new research project targeting a new type of material, property, or relation must first undertake considerable effort to create a large, carefully annotated training set tailored for this new target, a task that often requires considerable in-depth domain knowledge.

The subfield of polymer science puts these problems in particularly stark relief. Polymers have their own unique nomenclature, as we explain below, and thus annotated datasets created for general chemistry are of little value. Scientists and engineers lack access to a freely available large database of polymers and their properties. To address these challenges, we have previously designed polyNER [11, 12], a system for generating training data for scientific NER using semi-supervised and supervised learning. PolyNER uses NLP to produce sets of candidate entities, which experts approve or reject via a Web interface; the resulting labels are used to train context-based word vector classifiers. PolyNER's labels can also be used to train other machine learning models to leverage other features in addition to context, such as word morphology

Tchoua, Roselyne; Ajith, Aswathy; Hong, Zhi; Ward, Logan; Chard, Kyle; Audus, Debra; Patel, Shrayesh; de Pablo, Juan. "Active Learning Yields Better Training Data for Scientific Named Entity Recognition." Paper presented at 2019 IEEE 15th International Conference on e-Science (e-Science), San Diego, CA, US. September 24, 2019 - September 27, 2019.

to recognize target entities. The goal is to substitute the labor-intensive processes of assembling a large manually annotated corpus (and reduce costs) by using small numbers of carefully selected candidates to be labeled via focused expert input. We aim in our work to slash the expert time and effort required to achieve state-of-the-art NER performance in a new field.

In this paper, we seek to improve polyNER's performance by improving the labeling process and the classification of entity word vectors. Specifically, we address two challenges: (1) lack of (expensive) training data in some fields, including our own polymer science application which lacks free access to large polymer databases; and (2) the need for domain expert curators, which impedes the use of crowdsourcing platforms such as Amazon Mechanical Turk [13] or Figure8 (https://www.figure-eight.com/). In the initial labeling phase, we experiment with different *representative* (commonly used) entities to increase the fraction of target entities in the dataset to be labeled and bootstrap the context-based word vector classifiers. In the subsequent labeling and classification steps, we use active learning with maximum entropy uncertainty sampling and two different pools of unlabeled data to train classifiers, and compare their learning rate after five rounds. Using labels generated via active learning, we train word vector classifiers and achieve NER performance comparable to a modified version of a state-of-the art rule-based chemical entity extraction system, ChemDataExtractor (CDE) [9]. We have previously enhanced CDE with dictionary- and rule-based methods for identifying polymers [14]. Our system, however, took five hours of expert time to achieve this result.

The rest of this paper is as follows. In Section II, we motivate the need for identifying polymer names in text. We review semi-supervised methods for NLP systems in Section III. We describe the design and implementation of our active learning-based approach in Section IV. We evaluate our approach in Section V. We summarize and discuss future work in Section VI.

II. Motivation

Our work is generally motivated by the need to extract previously unmined scientific entities; our initial goal is to enable machine-learned extraction of polymer names. The challenges of polymer science NER are similar to those in biomedicine [2, 15]. Entities can be described with multiple referents (synonymy). Conversely, the same word may refer to different concepts depending on context (polysemy). For example, *polystyrene* is often referred to as *PS*, but *polystyrenes* can also be referred to as *GPPS*, *HIPS*, and *EPS*; combined with other monomers yielding *SBR*, *SBS*, and *ABS*; or used to describe polystyrene derivatives such as *PAMS*, *PMS*, and *PSS*. While standards for naming polymers exist (e.g., International Union of Pure and Applied Chemistry (IUPAC [16]) naming conventions), they are not always followed in practice [17]. Instead, polymer names may be reported as source-based names (based on the monomer name), structure-based names (based on the repeat unit), common names (requiring domain specific knowledge), trade names (based on the manufacturer), and names based on chemical groups within the polymer (requiring context to fully specify the chemistry), generating variability in naming conventions. Typographical variants (e.g., alternative uses of hyphens, brackets, spacing) and alternative component orders cause more variations between polymer names in the literature. The origin of these different naming conventions is linked to the desire for clarity within a journal article, coupled with the often-complicated monomeric structures [18].

In addition to challenges related to the makeup of the scientific entities, the scarcity of entities in scientific literature and the lack of training data also impede the use of machine learning based NER techniques. Considerable time and manual effort are involved in creating and maintaining the balanced CoNLL dataset for standard NER [19]. Example sentences from such corpora include one or more entities per sentence. In our attempt to recognize polymer names in full text documents, we face a very imbalanced dataset where most sentences do not contain a target entity, as there are only a handful of target entities per document. While there has been much interest in machine learned recognition of biochemical entities [7–9, 20], the successes that have been achieved have required much human effort to generate quality training data [2]. Previous work has also found that even state-of-the-art NER systems rarely perform well when applied to different domains [21]. Therefore, the problem of training machine learning models to recognize new scientific entities in a new field, such as polymer science, remains challenging.

III. Related Work

NER and other information extraction tasks rely on a large amount of training data, which are expensive to obtain. Weakly supervised learning methods work with much less training data and aim to address this challenge. They generally fall under two categories: semi-supervised learning and active learning [22]. The key difference between the two is that the former relies on approximately labeled data (as opposed to correctly labeled data for supervised learning) and the latter starts off with unlabeled data. Semi-supervised learning attempts to label data automatically by using prior knowledge and a set of labeled data. For example, it assumes that if x and y are similar, they probably have the same label (first cluster the whole dataset, then label each cluster with labeled data [23]). Active learning assumes there is a source of knowledge, such as a human expert, that can be queried to label a selected batch of unlabeled data.

A. Semi-supervised Approaches

Bootstrapping is a semi-supervised technique, which starts from a small set of seed relation instances and iteratively learns more relation instances and extraction patterns. Snowball [24] which improved the DIPRE system [25], used an intuitive idea to collect new entity relations using a set of seed entity pairs. In the DIPRE system, the intuitive assumption is that, given a few seed entity relations, the text between two known target entities in close proximity of each other describes and constitutes a *pattern* of the relation between the two. Since

Tchoua, Roselyne; Ajith, Aswathy; Hong, Zhi; Ward, Logan; Chard, Kyle; Audus, Debra; Patel, Shrayesh; de Pablo, Juan. "Active Learning Yields Better Training Data for Scientific Named Entity Recognition." Paper presented at 2019 IEEE 15th International Conference on e-Science (e-Science), San Diego, CA, US. September 24, 2019 - September 27, 2019.

that is not the case in practice, the system uses a limited set of regular expressions to limit useful patterns, hence decreasing the number of false positives. A key improvement of Snowball is that its patterns include named entity tags (PERSON, LOCATION, ORGANIZATION, etc.). Given a handful of seed tuples of ORGANIZATION and LOCATION, Snowball attempts to learn the relation *HeadquarteredIn* by assuming that each time the tuples appear in close proximity to each other, the text in between illustrates the desired relation. This text can then be used to discover new tuples, which can in turn be used as seeds for the next discovery round. Of course, organizations may be located but not headquartered in multiple cities; hence it is important to inspect the quality of extraction patterns to reduce noise in the generated output.

Distant supervision maps known entities and relations from a structured knowledge base onto unstructured text [26, 27]. With freely available structured knowledge base such as DBPedia [28] and Freebase [29], it is possible to leverage a large set of known entity pairs to generate training data. Over the past decade, probabilistic approaches have been proposed to allow automatic selection. For example, PaleoDeepDive [26], built upon DeepDive [27], automatically extracts paleontological data from text, tables, and figures in scientific publications. For good performance, such approaches often rely on and extend large databases: for example, PaleoDeepDive uses PaleoDB [30]. The system labels any entity pair that appears in the database as *True*. The user defines features (e.g. if a specific keyword appears between two entities, that pair a certain attribute is labeled *True*, but if the entity pairs are too far apart, another attribute is marked *False*), the system then uses statistical inference to determine the probability that each newly discovered pair of interest is *True*.

Data programming, as used in the Snorkel system [31], has users define *labeling functions* to provide labels for data subsets. Errors due to differences in accuracy and conflicts between labeling functions are addressed by learning and modeling the accuracies of the labeling functions. Under certain conditions, data programming achieves results on par with those of supervised learning methods. While writing concise scripts to define rules may seem to be a more reasonable task for annotators than exhaustively annotating text, it still requires expert guidance. In data programming as in bootstrapping and distant supervision it is important to evaluate the quality of functions and extraction patterns to decrease noisy patterns.

B. Active Learning

Active learning [22] assumes that gold standard labels for unlabeled instances can be obtained by querying an oracle (domain expert or source of knowledge). The goal of active learning is to decrease labeling costs by requesting a limited number of labels from the oracle, that have been deemed most valuable by the learner. Uncertainty sampling approaches define “valuable” data by measuring uncertainty in the predictions. For example, in the case of a single learner, querying predictions with maximum entropy in which the learner assigns all classes with equal probability [32] or predictions closest to the decision boundary in the case of support vector machine classifiers [33]. In the case of multiple learners, query-by-committee requests labels for unlabeled instances on which the learners disagree the most [34]. Uncertainty sampling and query-by-committee are representative approaches based on informativeness, where informativeness measures how well an unlabeled instance helps reduce the uncertainty. Another selection criterion addresses representativeness, which measures how well an instance helps represent the structure of input patterns; in this case selection is made by querying data from unlabeled clusters of data [35, 36].

C. CDE and CDE+

We introduce briefly ChemDataExtractor (CDE), a state-of-the-art chemical NLP tool that combines a dictionary, expert-created rules, and machine learning algorithms. CDE was trained on the CHEMDNER corpus: a collection of 10 000 PubMed abstracts with 84 355 chemical entity mentions labeled manually by expert chemistry literature curators, following annotation guidelines specifically defined for this task [2]. In previous work, we modified CDE with manually defined polymer identification rules [14], creating what we term here CDE+. We compare our methods against both CDE and CDE+ in Section V.

D. Placing polyNER in Context

PolyNER has in common with prior work that it combines semi-supervised and active learning [35, 37]. The initial sampling of unlabeled data is similar to bootstrapping but applied to named entities rather than entity pairs. The initial batch of labels contains strings deemed *similar* to a few seed entities, with similarity determined by using word representations and vector distance measures. As this approach is approximate and is likely to include errors (words that have similar context to entities but are not actual entities), the next phase is expert labeling. Subsequent batches of labels are obtained via active learning and used to train a context-aware word vector classifier in the last phase. PolyNER can be used in a way that complements other scientific NER approaches. For example, it could be used as a scientific entity tagger (i.e., recognizer) to be used with data programming to extract polymer properties.

IV. Design and Implementation

As previously mentioned, our main goal in designing polyNER is to slash labeling costs by reducing the time and effort spent by experts to generate training data. Rather than labeling entire documents and phrases, annotators label proposed candidate entities to be classified. Earlier results show that with two hours of labeling we can achieve precision or recall (but not both) on par with state-of-the-art domain specific software, by selecting an ensemble of classifiers for discrimination [12].

Here, we refine polyNER components and incorporate active learning with different sampling strategies in order to further improve performance. PolyNER uses word representations and minimal domain knowledge (a few seed entities) to produce a small set of candidates for expert labeling; labeled candidates are then used to train named entity word vector classifiers. We integrate an active learning loop into polyNER's architecture to incrementally improve classifier performance.

In order to explore whether the use of word vector coordinates as features can accelerate the learning process, we define and compare three alternative sampling strategies: a random strategy that we use as a baseline, and two NLP-based filtering methods. We also apply these methods against two different candidate pools, one set of unlabeled nouns and another set of approximately labeled nouns deemed *similar* to commonly used known entities from our corpus. We describe our sampling strategies and approximate labeling in more details in this section. The general architecture of polyNER is illustrated in Figure 1. We also describe the labeling process, and the training and testing configuration for our word vector classifiers in Section IV-F

A. Computing Word Embeddings

A word embedding method maps each word in a document to a vector in an n-dimensional real vector space that represents the linguistic context in which the word appears. This mapping may be based, for example, on co-occurrence frequencies of words. We can then determine the similarity between two words by computing the distance between their corresponding vectors in the feature space.

We use Word2Vec, a recent, light-weight and easy-to-use implementation of context-based vector representations [38, 39]. Specifically, we use the Gensim continuous bag-of-words (CBOW) implementation of the Word2Vec algorithm [40] to generate vectors. Prior parameter tuning indicated that window size and vector size did not have a significant impact on the yield of polymers (less than 5%) on initial bootstrapping (see Section IV-C below). Nevertheless, for slightly higher yields, we set the Word2Vec `size` parameter to 100 and the `window_size` to 2; where `size` is the size of the vector, and `window_size` is an adjustable window of surrounding context word used to compute each embedding.

B. NLP-Filtering

The NLP filtering preprocessing step removes strings that are unlikely to be polymer referents. Hypothesizing that names of scientific entities will not, in general, be English vocabulary words, we also remove words found in the SpaCy dictionaries of commonly used English words [41]. We manually remove common polymer names, such as polystyrene and polyethylene, from the dictionaries. We use SpaCy's part-of-speech tagging functionality to remove non-nouns. We also eliminate strings that represent numbers (including numbers followed by common units) and remove extraneous characters from the beginning and end of each candidates to filter out mispellings. Finally, we remove plurals (e.g., polyamides, polynorbornenes), as they can represent polymer family names. Note that these steps are generalizable and applicable to multiple science fields. We refer to the set of words that results when these filtering operations are applied to our corpus as the *NLP-filtered candidates*. This set is the output of step 1 in Figure 1.

C. Initial Bootstrapping and Labeling

NLP filtering reduces the number of entities to be considered, and increases the target vs. non-target entity ratio. However, there still remain a large pool of potential candidates from which entities are to be selected, of which, in our experience, roughly 5% are polymer names. In order to avoid presenting experts with mostly negative examples, hindering meaningful classification, we boost the number of polymer entities in the first batch of candidates to be annotated by selecting strings with low word vector distance (see Section IV-A) from a set of *seed entities*: words that are observed to occur frequently in a subset of publications, or that are suggested by experts. We discuss this distance metric in more detail below. Based on preliminary experiments, we set the size of each batch of strings to be labeled to 200, or about an hour of expert time. We then train the initial classifier on this bootstrap set, using 80% of the data for training and 20% for testing. We subsequently used three different sampling strategies for following classifications.

D. Sampling Strategies

We implement three sampling strategies, which we refer to as *Random*, *Uncertainty-Based Sampling (UBS)* and *Distance Uncertainty-Based Sampling (Distance UBS)*. We apply each of these strategies to our NLP-filtered candidates to determine which candidates to present to experts for labeling.

1) Random: Here, we randomly select 200 of the NLP-filtered candidates.

2) Uncertainty-Based Sampling (UBS): Our second strategy applies maximum entropy sampling to the NLP-filtered candidates. As previously mentioned, maximum entropy selection is an uncertainty sampling method that identifies data points for which a classifier predicts outcomes that lie near the decision boundary between classes. Thus, when predicting whether or not a word vector represents a polymer, maximum entropy arises when the classifier assigns equal probability to the polymer and not-polymer cases. As we have two classes, this equal probability is 0.5. We use the classifier to obtain a probability p for each NLP-filtered candidate. We select the 200 entries for which p is closest to 0.5 as our sample.

3) Uncertainty-Based Sampling with Distance Ranking (Distance UBS): Our third strategy is identical to UBS, except that it works with just a subset of the NLP-filtered candidates, namely the 10 000 that are closest to a set of seed entities. The intuition here is that a candidate is more likely to be a target referent (a name, acronym, synonym, etc.) if it used in a similar context. For example, the polymer name "polystyrene" in a sentence "The melting point of polystyrene is ..." suggests that X may also be a polymer in the sentence "The melting point of X is ...".

We use the word embeddings introduced in Section IV-A to capture this notion of context, and vector distance between word vectors as a measure of similarity. Whether or not this

Tchoua, Roselyne; Ajith, Aswathy; Hong, Zhi; Ward, Logan; Chard, Kyle; Audus, Debra; Patel, Shrayesh; de Pablo, Juan. "Active Learning Yields Better Training Data for Scientific Named Entity Recognition." Paper presented at 2019 IEEE 15th International Conference on e-Science (e-Science), San Diego, CA, US. September 24, 2019 - September 27, 2019.

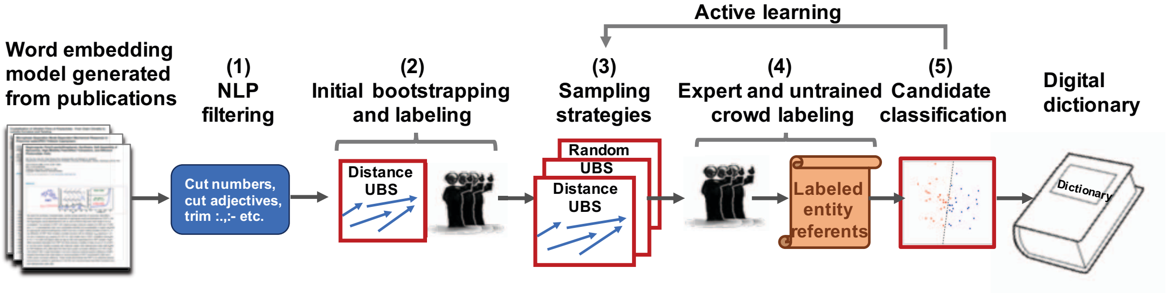

Active learning
Word embedding model generated from publications
(1) NLP filtering
Cut numbers, cut adjectives, trim :.,;- etc.
(2) Initial bootstrapping and labeling
Distance UBS
(3) Sampling strategies
Random
UBS
Distance UBS
(4) Expert and untrained crowd labeling
Labeled entity referents
(5) Candidate classification
Digital dictionary
Dictionary

Fig. 1: PolyNER system showing the different phases of polyNER including the NLP-filtering step, the initial bootstrapping and labeling phase as well as the newly integrated active learning loop to classify scientific named entities.

approach works in practice will depend on whether polymer names are in fact used in consistent contexts as captured by our word embedding vectors.

We can then determine, for each NLP-filtered word, the extent to which it occurs in a similar context to the seed entities, by computing the similarities between the word's vector and those for our seed entities. As we explain in Section V-B, we experiment with one and more seed entities; when dealing with multiple seed entities, we use the lowest distance score for ranking candidates.

E. Bootstrapping

The UBS and Distance UBS sampling methods use a classifier to determine which NLP-filtered entities should be chosen next for expert labeling. This classifier must be trained, and thus we need an initial set of entities to bootstrap this process. We could choose NLP-filtered entities at random for initial labeling, but that choice is unlikely to perform well due to the low proportion of polymers (just 5%) in the NLP-filtered corpus. Instead, therefore, we create an initial bootstrap set comprising the 200 NLP-filtered entities that are closest, in word distance vectors, to a set of seed entities. We then train the initial classifier on this bootstrap set, using 80% of the data for training and 20% for testing.

F. Active Learning Loop

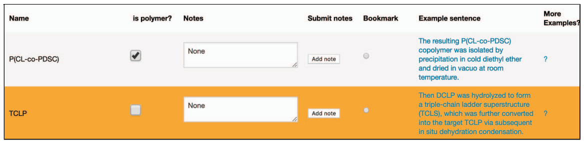
We now discuss our active learning process. As discussed in Section III-B, the basic idea here is that we repeatedly select a set of 200 candidate entities (a "sample") for expert labeling, based on what we have learned from previously labeled entities. We run this process independently with the Random, UBS, and Distance UBS sampling strategies, in order to compare their performance.

1) Use and Evaluation of Classifiers: Not specified in Section IV-D is the nature of the classifier that the UBS and Distance UBS strategies use to estimate the probability of each entity in the NLP-filtered corpus (or, for Distance UBS, the 10 000-entity subset of the NLP-filtered corpus) being a polymer. (The Random strategy does not use a classifier for sampling, as it selects candidates at random.) As we have no prior knowledge of the distribution of target entities in the vector space, we consider seven distinct classifiers in each round of the active learning process: the scikit-Learn [42] implementations of Decision Tree, Gradient Boosting, K-Nearest Neighbor (KNN), Logistic Regression, Linear Support Vector Machine, Naive Bayes, and Random Forest. In each case, we use the word embedding for each string as input features. In each round i, we train these seven classifiers on the sample data gathered in rounds j, $j<i$, and then use the classifier with the highest recall to determine the p scores that UBS and Distance UBS use when assembling their 200-entity sample in that step. (We use recall, or retrieving a maximum of targets, as a measure of performance, because we want to favor extracting a higher number of targets, potentially requiring additional curation, over obtaining fewer correct targets.)

The 200 entities in the new sample are passed to experts for annotation, and the annotated data are added to the set of training data used in the next learning round.

2) Expert (and Non-expert) Labeling: We engage two domain experts to annotate the candidates generated by the *UBS* and *Distance UBS* sampling strategies. Each expert annotates one strategy; we also perform crosschecking for 10% of the first batch of labels, to get a measure of agreement between experts, with results reported in Section V-C. Experts use a web interface (see Figure 2) to approve or reject candidates, a task that is far more efficient than reading and annotating words in text. The interface provides example sentences as context for ambiguous candidates and allows the expert to access the publication(s) in which a particular candidate appears.

We aim to reduce the amount of costly expert time used to obtain labels. Therefore, for our baseline of randomly sampled NLP-filtered nouns, we experiment with a two-phase review process. Tokenization is one of the largest sources of error for scientific entities such as polymers, which contain characters such as ':', '(', '–', ',' etc. Tokenization can also generate incoherent tokens from text, equations, captions, etc. Such obvious non-candidates can be fairly easily detected by non-experts. For example, an untrained human annotator may be able to recognize that '$d\Sigma/d\Omega)(Q$' is not a polymer name, and thus save time for the experts. Hence, we assigned two graduate student labelers to curate the candidates generated by the random sampling strategy, which are less likely to contain



Name	is polymer?	Notes	Submit notes	Bookmark	Example sentence	More Examples?
P(CL-co-PDSC)	☑	None	Add note	○	The resulting P(CL-co-PDSC) copolymer was isolated by precipitation in cold diethyl ether and dried in vacuo at room temperature.	?
TCLP	□	None	Add note	○	Then DCLP was hydrolyzed to form a triple-chain ladder superstructure (TCLS), which was further converted into the target TCLP via subsequent in situ dehydration condensation.	?

Fig. 2: Web interface for expert review of candidates. The expert indicates whether the name (column 1) is a polymer (tickbox in column 2), providing notes if desired (column 3). Clicking on "?" delivers up to 25 more example sentences.

target entities. We asked these untrained labelers to reject obvious non-candidates via the previously mentioned web interface. Our experts then reviewed the remaining candidates, indicating for each whether it is in fact a polymer referent.

V. Evaluation

We first report on a study in which we evaluate the generation of candidate entities using vector distances from representative (frequently used) entities. We then discuss the results of initial classification and subsequent four rounds of active learning using multiple word vector classifiers and our three sampling strategies: random, UBS, and Distance UBS. Finally, we experiment with word representations enhanced with character-level information using FastText [43, 44].

In this work, we evaluate extraction accuracy in terms of precision and recall. *Precision* refers to the fraction of predicted positives that are labeled correctly and *recall* to the fraction of actual positives that are labeled correctly.

A. Dataset

We reuse a corpus of 1690 full-text publications in HTML format from *Macromolecules*, a relevant journal in polymer science [12]. These documents comprise 381 947 sentences and 9 229 417 (253 195 unique) words or "tokens," of which 23 205 pass the NLP filter of Section IV-B. From this corpus, we set aside a test set of 100 documents with 22 664 sentences and 508 391 (36 293 unique) tokens, of which 9656 pass the NLP filter. Six experts identified all one-word polymer names in this test set, a process that produced 467 unique one-word polymer names. We use these 467 names as a gold standard in subsequent subsections; we automatically label all NLP-filtered strings from the 100-document test set using these manually extracted names.

B. Seed Entities

Recall from Section IV-A that polyNER uses the Word2Vec word embedding tool to compute a word vector for each word. In order to maximize the number of actual entities in the dataset—and the ratio of target to non-target entities—in the initial set of labels, we explore how the choice of seed entities impact the number of target entities retrieved. While we cannot expect meaningful classification using only positive examples, given the imbalance in the whole dataset, we aim to select the Word2Vec parameters that yield the highest ratio of polymers in this initial batch of candidates.

In the experiments that follow, we use the 467 gold standard polymer names identified by experts in our 100-document test set to evaluate performance with different seed entities. Specifically, for each choice of seed entities that we want to evaluate, we determine the 10 000 NLP-filtered words with vectors closest to the seed entity vectors, and report what fraction of the 467 gold standard names are included in that 10 000. We use lower-case exact string matching between the gold standard polymer names and the proposed distance candidate strings to determine if a candidate is a polymer.

In previous work using the same corpus [45, 46], we built a dictionary of polymer names by using a rule-based approach and aggregating synonyms across ChemDataExtractor records. (A record consists of all information found about a chemical entity in a document.) Here we use this dictionary to identify the 10 most frequently occurring polymers in our corpus and their acronyms. We assume that frequent polymers provide a large number of sentences that illustrate context in which polymers are commonly used. Hence, we first test the most frequent, the three most frequent, and the ten most frequent polymers as seed entities. We also experiment with including and excluding their acronyms as additional seed entities. (Note that this modest set of 1, 3 and 10 seed entities could also be suggested by an expert.)

Rows 1–6 of Table I shows the results for these experiments. When using *polystyrene* (the most frequently used name) as a seed entity, the candidates contained 33.6% of the 467 gold standard polymers. We note a 2% increase in the fraction of polymers retrieved when using both *polystyrene* and *PS*, when compared to using *polystyrene* alone. The fraction of polymers increases by 10% when we use three representative entities (the three most frequent polymers in our datasets are *polystyrene*, *poly(methyl methacrylate)*, and *polyethylene*), but by less than 1% when using 10 instead of three entities. These results suggest that there is little value to using more than a few seed entities.

TABLE I: Fraction of gold standard polymer names in the 10 000 entities that are closest, by word vector distance, to various sets of seed entities.

#	Seed entities	Fraction of polymers extracted
1	Polystyrene	35.6%
2	Polystyrene, with acronym PS	37.7%
3	Three most frequent polymer names	46.9%
4	Three most frequent polymer names, with acronyms	48.0%
5	10 most frequent polymer names	46.5%
6	10 most frequent polymer names, with acronyms	48.4%
7	χDB polymer names	46.7%
8	crowDB polymer names	36.4%

To further explore whether using larger numbers of seed entities may increase the fraction of polymers retrieved, we conducted a second set of experiments. We have built a small database of polymer properties (χDB) in previous work [45, 46]. Our corpus of 1690 publications included 111 out of 175 χDB polymers. We also scraped CrowDB, which lists some polymers and their properties at http://polymerdatabase.com/

Tchoua, Roselyne; Ajith, Aswathy; Hong, Zhi; Ward, Logan; Chard, Kyle; Audus, Debra; Patel, Shrayesh; de Pablo, Juan. "Active Learning Yields Better Training Data for Scientific Named Entity Recognition." Paper presented at 2019 IEEE 15th International Conference on e-Science (e-Science), San Diego, CA, US. September 24, 2019 - September 27, 2019.

for polymer names; 32 out of 295 scraped polymer names were found in our corpus. We measure how many of our gold standard polymers are identified when these 111 and 32 polymer names are used as seed entities, with results shown in rows seven and eight of Table I. These results confirm that using more entities does not increase the yield of polymers. Thus, in all subsequent experiments, we use the three most frequent polymers and their acronyms as seeds.

C. Labeling

We conduct some experiments to estimate labeling time. We ask two polymer scientists to label 200 candidates from a subset of our corpus. One expert reports 30 minutes, the other 45 minutes. We overestimate the time to label 200 candidates at one hour of expert time. Based on user feedback, we also improve the labeling Web interface after the above mentioned test rounds to further facilitate and speed up labeling. For example, we increase the number of example sentences available to provide context, to decrease the number of occurences in which experts have to look up original publications for candidates. We also increase the size of checkmarks to make it easier to reject erroneous candidates. In the initial labeling round, we perform crosschecking for 10% of the batch of labels. We confirm agreement between labels for all but one of 20: an agreement rate of 95%.

D. The Initial Classifier

Recall from Section IV-E that we use an initial set of 200 NLP-filtered entities that are close to seed entities to bootstrap our labeling process. Once those data are labeled by our experts, we use 80% to train a KNN classifier, validating on the remaining 20%. We then test this *initial classifier* against all 9656 NLP-filtered candidates in the 100-document test set, which as noted in Section V-A contain 467 polymers. Results are shown in Figure 3. Its Receiver Operating Characteristic (ROC) curve shows better-than-random behavior, with an area under curve (AUC) of 0.62. However, in our application, correctly predicting non-polymers is not as important as correctly identifying our targets. Therefore, we also plot in the Precision Recall (PR) curve to show the tradeoff between precision and recall. While the AUC for the initial classifier is above random performance (0.5) its PR curve shows poor precision, regardless of recall. In previous work [12], we found that we could achieve better performance with more labels (897), suggesting that a KNN classifier begins to learn with more data. However, 160 labels (80% of 200) is not yet enough.

E. Comparing Sampling Strategies

After the initial round of labeling, we experiment with the three sampling strategies described in Section IV: Random, UBS, and Distance UBS, performing four rounds of active learning for each. Results, in Table II, show a significant amount of fluctuation and no notable improvement in the first two rounds for any strategy: precision remains under the initial precision of 6.5% for all. However, in the third round, we observe increases in precision, an improvement that is sustained in the fourth round for UBS. Figure 4 shows ROC and PR curves for the three strategies after four rounds. The AUC for UBS is 0.74 and that of Distance UBS is 0.70. The PR curves for both are improved (lifting away from the lower left corner of the graph) over the first round, with active learning performing better with UBS than with Distance UBS. When tested against our gold standard of 467 one-word polymer names, the KNN classifier achieves 18.2% precision and 45.6% recall. We notice that even the random strategy PR curve is improved (away from the initial PR curve and close to the Distance UBS curve), indicating that the NLP-filtering alone is enough to enable the learning process after 1000 labels. We also note that KNN was most-often selected across strategies and iterations; likely due to its inherent nature to optimize locally and our direct focus on finding similarities between observations.

We conclude that while the sampling step helps ensure that the classes are balanced in the initial batch of labels, restricting ourselves to just distance candidates, as is done by Distance UBS, does not yield better results than using active learning with all NLP-filtered candidates (UBS). Intuitively, basic UBS can find *useful* instances (target and non-targets) to be labeled from the entire word embedding space, while examples from Distance UBS are clustered around the seed entities that may be collocated in that space.

TABLE II: Precision and recall relative to the gold standard for the initial classifier (round 0) and the classifiers trained also with the increased data obtained in each of four learning rounds, 1–4.

		Strategy		
Round	**Metric**	**Random**	**UBS**	**Distance UBS**
0	Precision	6.5%		
	Recall	19.1%		
1	Precision	3.8%	3.2%	5.3%
	Recall	0.29%	93.6%	56.8%
2	Precision	1.5%	3.8%	5.4%
	Recall	1.5%	46.4%	10.1%
3	Precision	6.0%	**21.2%**	3.9%
	Recall	44.6%	**40.0%**	84.3%
4	Precision	12.3%	**18.2%**	7.2%
	Recall	33.3%	**45.6%**	51.9%

We selected seed entities based on their frequency in our corpus. This observation suggests that we could also study how the choice of seed entities impact of the performance of the classifier during the active learning process. We revisit this concept of *diversity* of labels in the Section V-G. Note that with limited training data and based solely on context, the classifier retrieves 45.6% (more than one third) of the gold standard polymers with a precision of 18.2%, after five hours of expert labeling. For comparison, an attempt to extract polymer names using the rule: *if the name contains "poly" extract it as a polymer*, gets recall of 41% and precision of 34% on the same dataset. We conclude that with our relatively small and noisy dataset (based on context-only information from entire documents of unstructured and uncurated text), we are able to achieve close to rule-based performance, using active learning and little labeling.

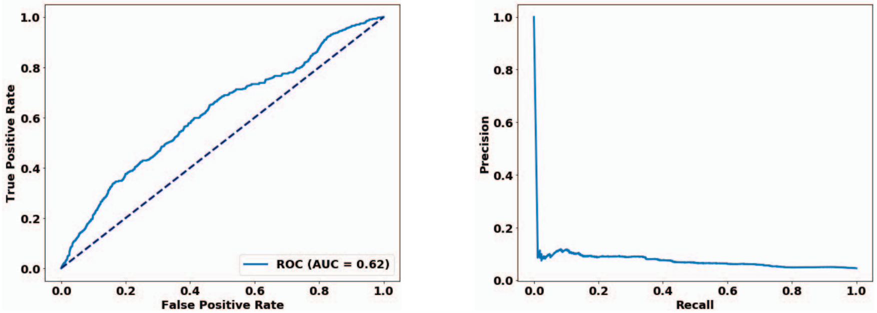


Fig. 3: ROC (left) and PR (right) curves for KNN model for the initial classifier. The PR curve shows that precision is low regardless of recall, indicating that we need more data.

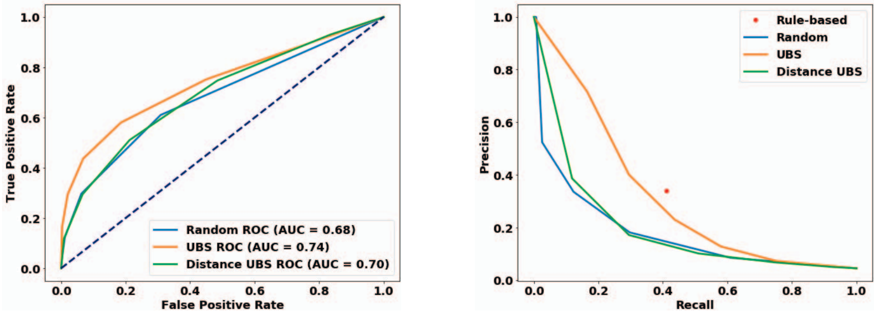


Fig. 4: ROC (left) and PR (right) curves after four rounds. The ROC curves for two active learning strategies, UBS and Distance UBS, show significant improvements, achieving AUCs of 0.74 and 0.70, respectively: significantly better than the 0.62 achieved by the initial classifier in Figure 3. Random achieves an AUC of 0.68. PR curves also show improvements relative to the initial classifier, with all three strategies lifting away from the bottom corner, indicating discriminative capacities. In both types of plots, UBS outperforms Distance UBS and approaches rule-based performance based solely on context information and under five hours of expert input.

F. Active Learning Labels + Character-Level Embeddings

After just 1000 labels, the context-based classifier using active learning applied to NLP-filtered candidates achieved performance comparable to rule-based performance, but not quite as good as the polymer-enhanced CDE+. With the goal of further improving polyNER performance, we experiment with the use of an alternative word embedding model, FastText, which uses word representations enriched with sub-word (character-level) information. Because FastText considers subword information as well as context, it can consider word morphology differences, such as prefixes and suffixes. Subword information is especially useful for words for which context information is lacking, as words can still be compared to morphologically similar existing words. We set the length of the sub-word used for comparison—FastText's `n_gram` parameter—to five characters, based on our intuition that many polymers begin with the prefixes "poly" or "poly(." Therefore, we generate a FastText word embedding model, and generate character-enhanced vectors for our UBS-labeled candidates.

Next, we train a KNN classifier using vectors for the candidates labeled through active learning from the NLP-filtered candidates (the active learning strategy identified as best-performing in Figure 4). We test the classifier against NLP-filtered nouns from our 100-document test set. KNN classifier performance improves when using these word vectors, achieving 29.7% precision and 81.9% recall, comparable to those achieved by CDE (see Section III-C). CDE's recall is high at 74.5%, but its precision for polymers is, as expected, low at 8.7%, as it does not incorporate polymer knowledge. In Figure 5, we show the PR curve for the FastText vector classifier and also results for the polymer-enhanced CDE+, which achieve 42.2% precision and 68.3% recall on the same test set. We achieve higher recall than CDE and CDE+ using labels from UBS and FastText vectors and in-between (higher than CDE and lower than CDE+) precision.

G. Discussion and Future Work

We have previously used seed entities to bootstrap classifiers of context-based word vectors. Using an ensemble of classifiers, polyNER allowed users to tradeoff precision and recall. In this attempt to improve performance, while efficiently using experts' time, we used active learning to obtain more labels. However, we do not observe an increase of precision over the 5% fraction of polymers until the third round of active learning with NLP-filtered words (800 labels). This suggests that our label batch size is lower than the minimum number of labels necessary to start the learning process. More detailed study of performance vs. label batch size will help pinpoint the appropriate number of labels and level of bootstrapping required for learning. We can also ask the following questions: Could we achieve CDE performance classifying only context-based (not character enhanced) representation? If so, how many more iterations would it take? If not, how much do we need to increase the size and/or improve our current corpus of 1690 documents? Corpora for NER tasks are curated and typically include sentences with one or more entities. We could investigate other semi-supervised method to eliminate negative sentences from our corpus (e.g. labeling sentences and classification of sentence vectors for example).

PolyNER achieves CDE performance using labels obtained via active learning (UBS sampling strategy) and FastText vectors. We attribute the increased performance to the character embedding enhancement, which not only recognizes "poly" (and yields more names based on this n-gram comparison), but also filters out more anomalous candidates (preceding or following polymer names) generated during tokenization and missed by the filtering steps, such as "A_mB_n" and "*Mw/Mn=1.36*." In other words, the classifiers of character and (context-based) word embedding vectors perform better than classifiers of only context-based word embedding. Given this result, one may wonder whether the active learning process itself could benefit from using this enhanced vector embedding. To determine whether this is the case, we repeated the active learning experiment using the entire corpus of NLP-filtered candidates and classifying FastText (enhanced) vectors instead of Gensim vectors at each round. However, the performance was worse than random.

These results suggest that character-level information enhances classifier performance only once a certain threshold of context information has been captured by the embedding. We explain this observation as follows. In FastText, the portion of the word embedding vector generated by using context varies depending on how much context is available in the entire corpus. For words deemed to have enough context, vectors do not include any character-level information. At the other extreme, for previously unseen words, the embedding is generated based solely on character n-gram information and comparison to other words in the corpus. During the active learning process, candidates to be labeled by experts are selected by using maximum-entropy-based uncertainty sampling: that is, words for which prediction probability is similar for target and non-target. Such candidates are more likely to lack context and thus have vectors that use mostly character-level information. As a result, the expert is often presented with nearly identical candidates (e.g., PS13k/PMMA12k, PS214k/PMMA12k, PS31.6k/PMMA12k), which hinders the learning process as these candidates are located in close vicinity in terms of the full (character and context) word embedding space. In other words, in this full space, while their uncertainty measure is comparable, these examples are not *diverse*, where diversity is a measure of the distance of the examples to each other or previously labeled instances [47]. One solution to explore in future work would be to impose a diversity constraint on the candidates, for example by using batch active learning [48]. This will be part of broader plan to study the effect of quantity (more documents) and quality (select paragraphs or sentences) of data on the word embedding model and the final results.

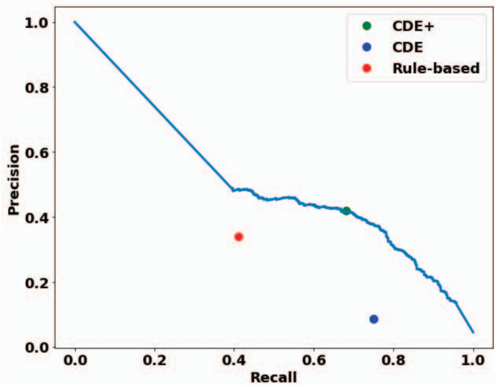


Fig. 5: PR curve for KNN model trained with active learning labels and word representations enriched with character-level information. Results for CDE+ are also shown. Note: PR curves are obtained by varying the threshold of probability that separates classes; straight lines occur when several points have similar probabilities and changing the threshold yields identical precision to recall ratio.

VI. Conclusion

A lack of expert-annotated training data impedes the adoption of machine learning techniques in certain scientific applications. PolyNER overcomes this challenge by using active learning to target expert input so that accurate scientific named entity recognition can be performed at low cost. We show that by using NLP techniques, we can bootstrap a word vector classifier of scientific entities. Using polyNER's labels and a classifier of character-enhanced word embedding vectors, we achieve performance comparable to a best-of-breed hybrid NER model (CDE+) that required much expert development. In contrast, polyNER was trained on data annotated using just five hours of expert time and a little untrained crowd input. Our work highlights the potential for using minimal labeled data and focused expert input to enable machine learning techniques for previously unmined scientific entities. We are currently exploring using polyNER-labeled data to annotate text for other NER approaches, such as bidirectional long short-term memory models. Our code and training dataset will soon be available on DLHub [49] for the public to use for training of machine learning models. Such resources can be used along with other databases and dictionaries such as PPPDB [1] and Khazana [50] for validation purposes. We will also formally explore the hybrid-computer partnership as an optimization problem. In other words, we will work on a more rigorous approach to automatic partitioning and assignment

[1] Polymer Property Predictor and Database: https://pppdb.uchicago.edu/

Tchoua, Roselyne; Ajith, Aswathy; Hong, Zhi; Ward, Logan; Chard, Kyle; Audus, Debra; Patel, Shrayesh; de Pablo, Juan. "Active Learning Yields Better Training Data for Scientific Named Entity Recognition." Paper presented at 2019 IEEE 15th International Conference on e-Science (e-Science), San Diego, CA, US. September 24, 2019 - September 27, 2019.

of extraction tasks in order to maximize the accuracy of extracted data while minimizing the time and cost of human involvement.

Acknowledgments

This work was supported in part by NIST contract 60NANB15D077, the Center for Hierarchical Materials Design, and DOE contract DE-AC02-06CH11357, and by computer resources provided by Jetstream [51]. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

References

[1] Y. Song *et al.*, "POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004," in *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004, pp. 100–103.

[2] M. Krallinger *et al.*, "CHEMDNER: The drugs and chemical names extraction challenge," *Journal of Cheminformatics*, vol. 7, no. 1, p. S1, 2015.

[3] J. J. de Pablo *et al.*, "The Materials Genome Initiative, the interplay of experiment, theory and computation," *Current Opinion in Solid State and Materials Science*, vol. 18, no. 2, pp. 99–117, 2014.

[4] R. Leaman and G. Gonzalez, "BANNER: An executable survey of advances in biomedical named entity recognition," in *Biocomputing*, 2008, pp. 652–663.

[5] Z. Zeng *et al.*, "Survey of natural language processing techniques in bioinformatics," *Computational and Mathematical Methods in Medicine*, 2015.

[6] L. Hawizy *et al.*, "ChemicalTagger: A tool for semantic text-mining in chemistry," *Journal of Cheminformatics*, vol. 3, no. 1, p. 17, 2011.

[7] T. Rocktäschel *et al.*, "ChemSpot: A hybrid system for chemical named entity recognition," *Bioinformatics*, vol. 28, no. 12, pp. 1633–1640, 2012.

[8] R. Leaman *et al.*, "tmChem: A high performance approach for chemical named entity recognition and normalization," *Journal of Cheminformatics*, vol. 7, no. 1, p. S3, 2015.

[9] M. C. Swain and J. M. Cole, "ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature," *Journal of Chemical Information and Modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[10] S. R. Young *et al.*, "Data mining for better material synthesis: The case of pulsed laser deposition of complex oxides," *Journal of Applied Physics*, vol. 123, no. 11, p. 115303, 2018.

[11] R. Tchoua *et al.*, "Towards hybrid human-machine scientific information extraction," in *2018 New York Scientific Data Summit (NYSDS)*. IEEE, 2018, pp. 1–3.

[12] R. B. Tchoua *et al.*, "Creating training data for scientific named entity recognition with minimal human effort," in *International Conference on Computational Science*, 2019.

[13] M. Buhrmester *et al.*, "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[14] R. B. Tchoua *et al.*, "Towards a hybrid human-computer scientific information extraction pipeline," in *13th International Conference on e-Science*, 2017, pp. 109–118.

[15] J.-D. Kim *et al.*, "Introduction to the bio-entity recognition task at JNLPBA," in *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 2004, pp. 70–75.

[16] R. C. Hiorns *et al.*, "A brief guide to polymer nomenclature," *Polymer*, vol. 54, no. 1, pp. 3–4, 2013.

[17] J. Tamames and A. Valencia, "The success (or not) of HUGO nomenclature," *Genome Biology*, vol. 7, no. 5, p. 402, 2006.

[18] D. J. Audus and J. J. de Pablo, "Polymer informatics: Opportunities and challenges," *ACS Macro Letters*, vol. 6, no. 10, pp. 1078–1082, 2017.

[19] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *7th Conference on Natural Language Learning*, 2003, pp. 142–147.

[20] D. M. Jessop *et al.*, "OSCAR4: A flexible architecture for chemical text-mining," *Journal of Cheminformatics*, vol. 3, no. 1, p. 41, 2011.

[21] M. Krallinger *et al.*, "Overview of the chemical compound and drug name recognition (CHEMDNER) task," in *BioCreative Challenge Evaluation Workshop*, vol. 2, 2013, p. 2.

[22] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, 2017.

[23] X. J. Zhu, "Semi-supervised learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep. 1530, 2005.

[24] E. Agichtein and L. Gravano, "Snowball: Extracting relations from large plain-text collections," in *5th ACM conference on Digital libraries*, 2000, pp. 85–94.

[25] S. Brin, "Extracting patterns and relations from the world wide web," in *International Workshop on The World Wide Web and Databases*, 1998, pp. 172–183.

[26] S. E. Peters *et al.*, "A machine reading system for assembling synthetic paleontological databases," *PLoS One*, vol. 9, no. 12, p. e113523, 2014.

[27] C. De Sa *et al.*, "DeepDive: Declarative knowledge base construction," *ACM SIGMOD Record*, vol. 45, no. 1, pp. 60–67, 2016.

[28] S. Auer *et al.*, "Dbpedia: A nucleus for a web of open data," in *The Semantic Beb*, 2007, pp. 722–735.

[29] K. Bollacker *et al.*, "Freebase: A collaboratively created graph database for structuring human knowledge," in *SIGMOD International Conference on Management of Data*, 2008, pp. 1247–1250.

[30] "Paleodb," http://paleodb.org, accessed March, 2019.

[31] A. J. Ratner *et al.*, "Data programming: Creating large training sets, quickly," in *Advances in Neural Information Processing Systems*, 2016, pp. 3567–3575.

[32] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *11th International Conference on Machine Learning*, 1994, pp. 148–156.

[33] C. Campbell *et al.*, "Query learning with large margin classifiers," in *17th International Conference on Machine Learning*, 2000, pp. 111–118.

[34] H. S. Seung *et al.*, "Query by committee," in *5th Annual Workshop on Computational Learning Theory*, 1992, pp. 287–294.

[35] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *21st International Conference on Machine Learning*, 2004, p. 79.

[36] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *25th International Conference on Machine Learning*, 2008, pp. 208–215.

[37] S. Basu *et al.*, "Active semi-supervision for pairwise constrained clustering," in *International Conference on Data Mining*, 2004, pp. 333–344.

[38] T. Mikolov *et al.*, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[39] ——, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[40] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Workshop on New Challenges for NLP Frameworks*, 2010.

[41] J. D. Choi *et al.*, "It depends: Dependency parser comparison using a web-based evaluation tool," in *53rd Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2015, pp. 387–396.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[43] P. Bojanowski *et al.*, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[44] A. Joulin *et al.*, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[45] R. B. Tchoua *et al.*, "A hybrid human-computer approach to the extraction of scientific facts from the literature," *Procedia Computer Science*, vol. 80, pp. 386–397, 2016.

[46] ——, "Blending education and polymer science: Semiautomated creation of a thermodynamic property database," *Journal of Chemical Education*, vol. 93, no. 9, pp. 1561–1568, 2016.

[47] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *20th International Conference on Machine Learning*, 2003, pp. 59–66.

[48] B. Settles, "Active learning literature survey," University of Wisconsin-Madison Department of Computer Sciences, Tech. Rep., 2009.

[49] R. Chard *et al.*, "Dlhub: Model and data serving for science," *arXiv preprint arXiv:1811.11213*, 2018.

[50] T. D. Huan *et al.*, "A polymer dataset for accelerated property prediction and design," *Scientific data*, vol. 3, p. 160012, 2016.

[51] C. A. Stewart *et al.*, "Jetstream: A self-provisoned, scalable science and engineering cloud environment," 2015.

NIST SP 1285
September 2022

Moisture Transfer in Commercial Buildings Due to Air Leakage: A New Feature in the Online Airtightness Savings Calculator

Som Shrestha[1]
Andre Desjarlais[1]
Laverne Dalgleish[2]
Lisa Ng[3]
Diana Hun[1]
Steven Emmerich[3]
Gina Accawi[1]

[1] *Building Technologies Research and Integration Center*
Oak Ridge National Laboratory, Oak Ridge, TN

[2] *Air Barrier Association of America*
Walpole, MA

[3] *Indoor Air Quality and Ventilation Group*
National Institute of Standards and Technology, Gaithersburg, MD

Content submitted to and published by:
Thermal Performance of the Exterior Envelopes of
Whole Buildings XIV International Conference

December 2019

U.S. Department of Commerce
Wilbur Ross, Secretary

National Institute of Standards and Technology
Walter G. Copan, Under Secretary of Commerce for Standards and Technology and Director

Shrestha, Som; Desjarlais, Andre; Dalgleish, Laverne; Ng, Lisa; Hun, Diana; Emmerich, Steven; Accawi, Gina. "Moisture Transfer in Commercial Buildings due to the Air Leakage: A New Feature in the Online Airtightness Savings Calculator." Paper presented at 2019 Buildings XIV International Conference, Clearwater, FL, US. December 09, 2019 - December 12, 2019.

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

Shrestha, Som; Desjarlais, Andre; Dalgleish, Laverne; Ng, Lisa; Hun, Diana; Emmerich, Steven; Accawi, Gina. "Moisture Transfer in Commercial Buildings due to the Air Leakage: A New Feature in the Online Airtightness Savings Calculator." Paper presented at 2019 Buildings XIV International Conference, Clearwater, FL, US. December 09, 2019 - December 12, 2019.

ABSTRACT

Air leakage through the building envelopes is responsible for a large amount of energy use. The US Department of Energy Windows and Building Envelope Research and Development Roadmap for Emerging Technologies states that, in 2010, air infiltration was responsible for 20% of primary energy consumption attributable to the fenestration and building envelopes of commercial buildings. Despite this fact, improving airtightness is not always recognized by commercial building owners, as they have been slow in acknowledging and diminishing the detrimental effects of air leakage on energy use, comfort, indoor air quality, and building material durability.

The design and construction industry would benefit from a credible, easy-to-use tool that estimates potential energy and financial savings in a standardized manner so designers and contractors can give building owners compelling reasons to invest in reducing air leakage. In 2016–2017, Oak Ridge National Laboratory, the National Institute of Standards and Technology, the Air Barrier Association of America, and the US-China Clean Energy Research Center for Building Energy Efficiency collaborated to develop an online calculator. This user-friendly calculator is free to the public and uses the simulation results of the whole building energy simulation tool EnergyPlus and the airflow simulation tool CONTAM. In 2018–2019, the calculator was expanded to add moisture transfer calculations given that air leakage through the building envelope can have a significant impact on moisture transfer and associated impacts. Four more commercial building types were also added to the existing database of three building types as part of this update. This paper describes the procedure used to calculate moisture transfer due to air leakage and provides examples that demonstrate the moisture transfer for each of the seven commercial building types that are currently part of the calculator.

INTRODUCTION

Envelope air leakage in commercial buildings in the United States accounts for about one quad (1 EJ) of primary energy annually (US Department of Energy [DOE] *Windows and Building Envelope Research and Development Roadmap for Emerging Technologies*), costing approximately $10 billion/year. The roadmap also states that computational tools are critically important for the design of commercial buildings with energy-efficient envelope technologies. As the thermal resistance of commercial building envelopes continues to improve, the relative contribution of air leakage to heating and cooling loads is increasing. As new technologies are developed, models and simulation tools must be updated to account for their performance. One impediment to the broader adoption of continuous air barrier systems into buildings is the lack of a simple, credible tool that can be employed by building architects, designers, and owners to accurately estimate the energy savings that can be expected if building envelope airtightness is improved.

Advances are needed in easy-to-use tools for determining the impact of air leakage to promote more energy-efficient and durable building envelope design. Oak Ridge National Laboratory (ORNL), the National Institute of Standards and Technology (NIST), the Air Barrier Association of America (ABAA),

Shrestha, Som; Desjarlais, Andre; Dalgleish, Laverne; Ng, Lisa; Hun, Diana; Emmerich, Steven; Accawi, Gina. "Moisture Transfer in Commercial Buildings due to the Air Leakage: A New Feature in the Online Airtightness Savings Calculator." Paper presented at 2019 Buildings XIV International Conference, Clearwater, FL, US. December 09, 2019 - December 12, 2019.

and the US-China Clean Energy Research Center for Building Energy Efficiency (CERC BEE) partnered to develop an online energy savings calculator[1], hereinafter referred to as “the calculator” (Shrestha et al. 2016), which calculated the potential energy and costs savings estimates from building envelope air tightening for three building types in 62 cities. The updated calculator now also calculates the reduction in moisture transport from improvements in airtightness, based on pre-and post-retrofit air leakage rates for commercial buildings. The updated calculator also includes four more building types than the original calculator. This article describes the need to add moisture transfer to the calculator, the buildings included in the calculator, the calculator itself, and the procedure used to calculate moisture transfer due to air leakage that is included in the updated calculator.

PREVIOUS STUDIES

Although air leakage through building envelopes has long been recognized as a major contributor to heating and cooling loads, methods that estimate the effects of air leakage on energy consumption vary (Crawley et al. 2008; Goel et al. 2014; Gowri et al. 2009; Ng et al. 2012). Air leakage has also been identified as a significant cause of moisture damage in lightweight constructions as demonstrated by the following studies. Belleudy et al. (2015) state that airtightness has become a considerable challenge over the past few decades in creating low-energy and durable buildings. The study assesses the impact of an airtightness defect on the hygrothermal field in a ceiling section insulated with loose-fill cellulose and separating a heated indoor space from an unheated attic. The experimental and simulation studies were conducted with and without air leakage, and the results show unambiguously that even relatively limited airflow through construction elements has a substantial impact on hygrothermal fields within the building envelope. Janssen and Hens (2003) identify air leakage as a significant cause of condensation problems in lightweight roof systems in cold climates. The study points out a need to incorporate airtightness requirements in building codes and to develop and certify adequate air barrier systems. Tenwolde and Rose (1996) conclude air leakage is one of the primary moisture sources that may increase the risk of moisture problems in wood frame walls. Similarly, Armstrong et al. (2010) identify air leakage as a potential cause of moisture damage in conventional light wood frame walls with exterior insulation that has low water vapor permeability.

PROTOTYPE BUILDING MODELS USED IN THE CALCULATOR

The calculator (described in the next section) uses DOE commercial prototype building models (DOE 2019) developed by DOE in EnergyPlus[2] (Deru et al. 2011) as a standardized baseline for energy savings calculations, and in part to support the development of ANSI/ASHRAE/IES Standard 90.1, *Energy Standard for Buildings except Low-Rise Residential Buildings*. The envelope assembly and heating, ventilating, and air conditioning (HVAC) units for each of the prototypes vary based on geographical location and the version of ASHRAE Standard 90.1 with which the prototype building was modeled to comply. A scorecard of each prototype building model, provided by DOE (DOE 2019), summarizes the building descriptions, thermal zone internal loads, schedules, and other key modeling input information.

The prototype models cover 16 commercial building types that represent about 80% of commercial

[1] https://airleakage-calc.ornl.gov/#/

[2] https://energyplus.net/

buildings in the United States in 17 climate locations defined in ASHRAE/IES Standard 90.1-2013 (ASHRAE 2013). Figure 1 shows the prototype buildings as a percentage of total US commercial building floor space. These are depicted in Figure 1 by a solid green-colored bar and represent over 55% of US commercial floor space. Figure 2 shows the renderings of the seven commercial prototype building models available in the calculator. Table 1 shows the floor area, number of floors, five-sided envelope area (exposed to ambient conditions), and six-sided envelope area of the 16 commercial prototype building models.

The variables defined in these models include building envelope components, HVAC equipment types and efficiency, and occupancy schedules. As ASHRAE Standard 90.1 evolves, Pacific Northwest National Laboratory modifies these models with input from ASHRAE Standing Standard Project Committee 90.1 members and other building industry experts. Features of the building models and a detailed description of their development are provided by Goel et al. (2014) and the Building Energy Codes Program website (DOE 2019).

The selection of the included cities was based on consideration of the major metropolitan areas throughout the United States; therefore, not every state or province is represented. If a specific city of interest does not appear on the list, it is recommended that the user select a city that has similar meteorological conditions (wind, temperature, solar radiation, and rain). This is not always the city geographically closest to the target city. For example, Minneapolis (MN) is in climate zone 6A and commercial prototype building models are not available for Minneapolis; so, the models for Burlington (the representative city for climate zone 6A), along with the weather file for Minneapolis, were used to run simulations for Minneapolis. Models that represent typical commercial buildings in Canada and China are not available in the public domain; therefore, the DOE prototype building models were used in these two countries. US cities where the commercial prototype building models were available, and where the climate matched the five cities in Canada and China, were used to model the buildings in those countries.

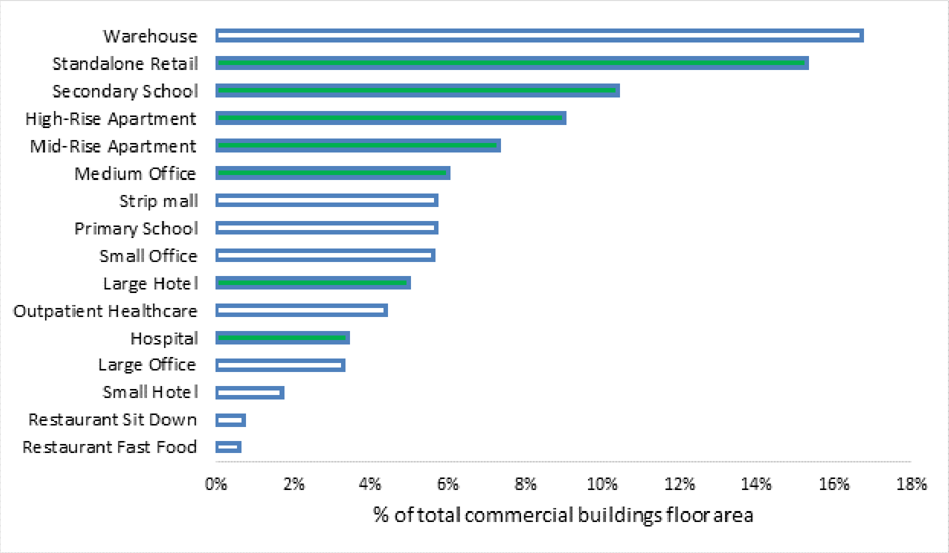
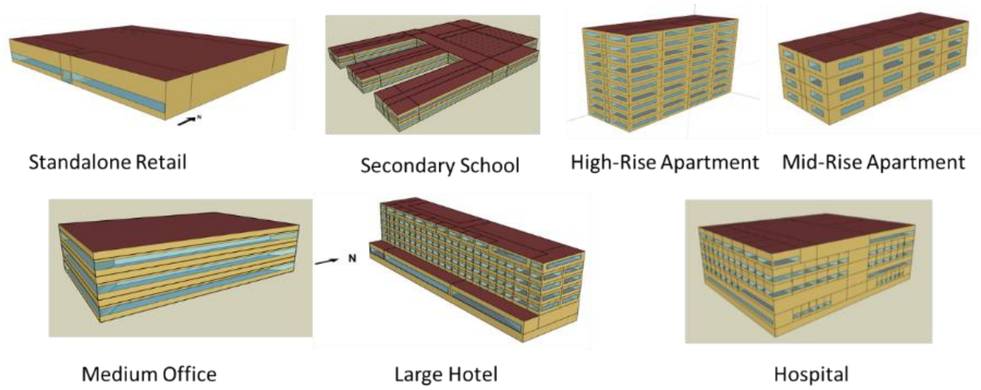


Figure 1. Prototype buildings as a percentage of total US commercial building floor space.

Table 1. Some Details of the Prototype Building Models

Building	Floor Area, m^2	Number of floors	5-sided envelope area, m^2	6-sided envelope area, m^2
Standalone Retail	2,294	1	3,471	5,765
Mid-Rise Apartment	3,131	4	2,326	3,109
Medium Office	4,980	3	3,640	5,301
High-Rise Apartment	7,837	10	4,639	5,422
Hospital	22,428	5	9,089	12,827
Large Hotel	11,346	6 + basement	6,005	7,984
Secondary School	19,593	2	17,871	29,774
Small Hotel	4,013	4	2,697	3,701
Large Office	46,321	12 + basement	15,158	18,726
Small Office	511	1	880	1,392
Outpatient Healthcare	3,804	3	2,938	4,322
Restaurant Fast Food	232	1	445	677
Restaurant Sit Down	511	1	845	1,356
Strip Mall	2,090	1	3,274	5,365
Primary School	6,871	1	9,384	16,256
Warehouse	4,598	1	7,095	11,694

Note: 1 m^2 = 10.76 ft^2



Standalone Retail

Secondary School

High-Rise Apartment

Mid-Rise Apartment

Medium Office

Large Hotel

Hospital

Figure 2. Renderings of the seven commercial prototype building models available in the calculator.

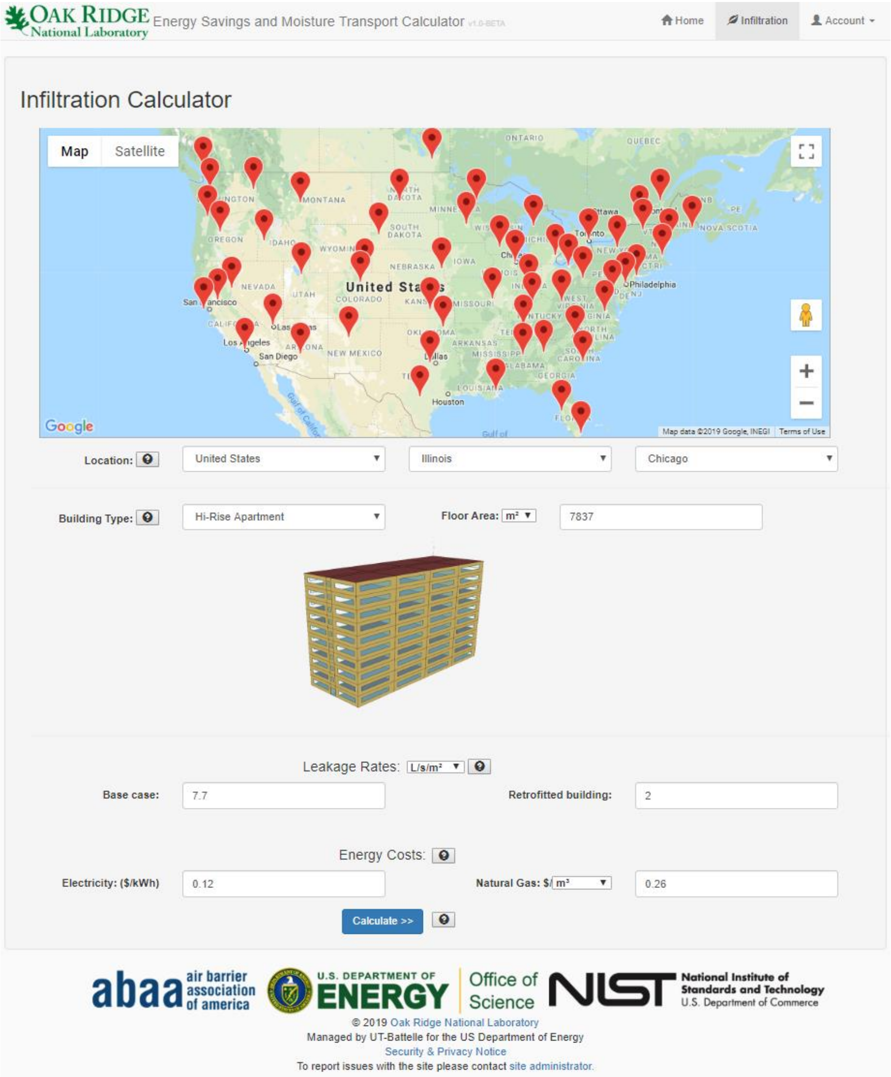
THE CALCULATOR

The calculator uses a database of pre-run simulation results from DOE's whole-building energy simulation software EnergyPlus for the DOE commercial prototype buildings. The main difference between the online calculator and the procedure followed in the DOE prototypes to account for infiltration (or "air leakage") is that the calculator uses CONTAM-calculated air leakage rates as inputs into EnergyPlus, whereas the prototype models make more simplified assumptions regarding air leakage. CONTAM (Dols and Polidoro 2016) is a multizone airflow and contaminant transport analysis software developed at NIST. This software considers multiple factors, such as weather conditions, envelope airtightness, and HVAC system operation, to calculate air leakage rates through building envelopes. The CONTAM-calculated hourly air leakage rates are converted into the format required by EnergyPlus using the CONTAM Results Export Tool (Polidoro 2016). EnergyPlus is then used to calculate the effects of air leakage on energy consumption and moisture transport.

Typical energy simulations tend to simplify their analyses by assuming constant air leakage rates or using simplified algorithms that can lead to less accurate energy usage estimates. Ng et al. 2018 estimate that these simplifications in the EnergyPlus models for the prototype commercial buildings lead to underestimations of average electrical and gas use for heating and cooling. Shrestha et al. 2016 show that the discrepancy in the predicted cost savings could be as high as 40%.

The moisture transport calculation in the calculator (as described below) computes the total amount of moisture that could be transported through the building envelope as a result of air leakage, assuming no loss or gain while traveling through the building envelope. It is a measure of the potential moisture source but does not quantify how much moisture is accumulating in the wall assembly. The hypothesis is that more moisture transported through the wall, the higher the likelihood it will create a durability issue and other moisture-related problems such as mold growth.

Figure 3 shows the input page of the calculator. The user input parameters are location (country, state, city), building type (one of the seven building types from the drop-down menu), floor area, building envelope leakage rate at 75 Pa (0.3-inch water column), "base case", and reduced air leakage after improving envelope airtightness, the "retrofitted building", and unit energy cost (electricity and natural gas). Location can be selected either by using drop-down menus or by using the map. Cities available in the database are highlighted with red flags on the map. Once the building type is selected, the default footprint of the corresponding prototype building is displayed. However, the user can change the floor area to their building footprint. The calculator prorates the energy savings and moisture transport results based on the floor area input by the user. Descriptions of each input variable and recommendations can be obtained by pressing the help button. The calculator allows data input in either SI or imperial units.



Infiltration Calculator

Location: United States | Illinois | Chicago

Building Type: Hi-Rise Apartment

Floor Area: m² 7837

Leakage Rates: L/s/m²

Base case: 7.7

Retrofitted building: 2

Energy Costs:

Electricity: ($/kWh) 0.12

Natural Gas: $/m³ 0.26

abaa air barrier association of america | U.S. DEPARTMENT OF ENERGY | Office of Science | NIST National Institute of Standards and Technology U.S. Department of Commerce

© 2019 Oak Ridge National Laboratory
Managed by UT-Battelle for the US Department of Energy
Security & Privacy Notice
To report issues with the site please contact site administrator.

Figure 3. Input page for the Energy Savings and Moisture Transfer Calculator.

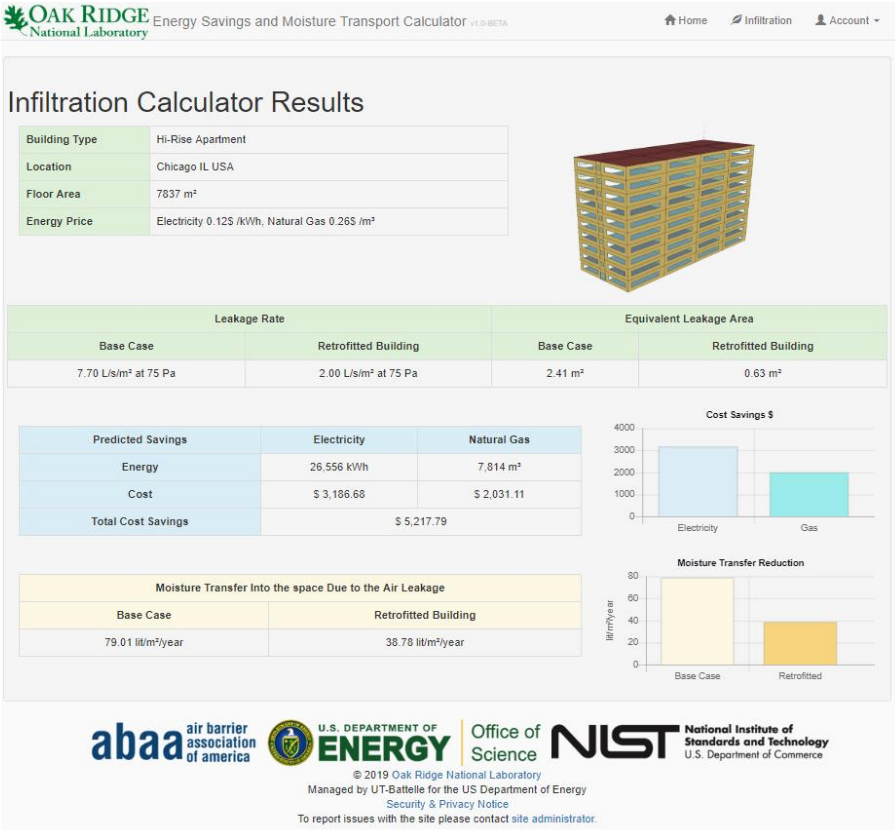
The output screen is shown in Figure 4. A summary of the user inputs is posted at the top of the page. The calculator determines the equivalent leakage area at 4 Pa (0.3-inch water column) (ELA) for the base case and the retrofitted building. This is calculated using Chapter 16 in the *ASHRAE Handbook—Fundamentals 2017*. When the ELA is calculated, all openings in the building shell are combined into an overall opening area and discharge coefficient for the building. The ELA of a building is, therefore, the area of an orifice with an assumed value of discharge coefficient that would produce the same amount of leakage as the building envelope at the reference pressure. The calculator also calculates the amount of energy saved (electricity and natural gas) and the cost savings in the currency of the country. Finally, the calculator computes the total amount of moisture that could be transported through the wall for both the base and retrofitted cases.

Table 2 compares the moisture transfer at an envelope leakage rate of 7.7 L/(s•m^2) (1.5 cubic feet per minute (CFM)/ft^2) and 1.25 L/(s•m^2) (0.25 CFM/ft^2) at 75 Pa (0.3-inch water column) for the seven building types in Chicago, including the percent reduction in the last column as a result of the increase in building envelope airtightness.

Table 2. Moisture Transfer for all Buildings at Two Different Airtightness levels at Chicago, IL

Building Type	Moisture Transfer, kg/(m^2•year)		Reduction in Moisture Transfer, kg/(m^2•year) (% reduction)
	At 7.7 L/(s•m^2)	At 1.25 L/(s•m^2)	
Standalone Retail	105.7	11.6	94.2 (89)
Mid-Rise Apartment	90.8	14.7	76.1 (84)
Medium Office	103.9	10.5	93.4 (90)
High-Rise Apartment	79.0	27.8	51.2 (65)
Hospital	73.1	12.1	61.0 (83)
Large Hotel	112.9	60.0	52.9 (47)
Secondary School	153.2	40.1	113.1 (74)

Note: 1 kg/(m^2•year) = 0.205 lb./(ft^2•year), 1 L/(s•m^2) = 0.2 CFM/ft^2.



OAK RIDGE National Laboratory
Energy Savings and Moisture Transport Calculator v1.0-BETA
Home
Infiltration
Account

Infiltration Calculator Results

Building Type	Hi-Rise Apartment
Location	Chicago IL USA
Floor Area	7837 m²
Energy Price	Electricity 0.12$ /kWh, Natural Gas 0.26$ /m³

Leakage Rate		Equivalent Leakage Area	
Base Case	Retrofitted Building	Base Case	Retrofitted Building
7.70 L/s/m² at 75 Pa	2.00 L/s/m² at 75 Pa	2.41 m²	0.63 m²

Predicted Savings	Electricity	Natural Gas
Energy	26,556 kWh	7,814 m³
Cost	$ 3,186.68	$ 2,031.11
Total Cost Savings	$ 5,217.79	

Moisture Transfer Into the space Due to the Air Leakage	
Base Case	Retrofitted Building
79.01 lit/m²/year	38.78 lit/m²/year

Moisture Transfer Reduction
lit/m²/year
80
60
40
20
0
Base Case
Retrofitted
abaa air barrier association of america
U.S. DEPARTMENT OF ENERGY
Office of Science
NIST National Institute of Standards and Technology U.S. Department of Commerce

© 2019 Oak Ridge National Laboratory
Managed by UT-Battelle for the US Department of Energy
Security & Privacy Notice
To report issues with the site please contact site administrator.

Figure 4. Output page for the Energy and Moisture Transfer Calculator.

EXAMPLE CALCULATIONS

This section presents example calculations for the DOE High-Rise Apartment prototype building model using the updated calculator. The relevant characteristics of this prototype model are based on ASHRAE 90.1-2013 and listed in Table 1. The scorecard for High-Rise Apartment building shows that this building was modeled with an air leakage rate of 1 L/(s•m^2) (0.2 CFM/ft^2) of exterior envelope area at 4.47 m/s (880 ft/min) wind speed. In the EnergyPlus model of this building, infiltration is modeled using the "ZoneInfiltration:DesignFlowRate" method that calculates infiltration using Eq. (1).

$$Infiltration = \left(I_{design}\right)\left(F_{Schedule}\right)\left[A + B\left|\left(T_{zone} - T_{odb}\right)\right| + C\left(WindSpeed\right) + D\left(WindSpeed^2\right)\right] , \qquad (1)$$

where

I_{design} = design infiltration volume flow rate normalized by exterior surface area, m^3/(s•m^2)

$F_{schedule}$ = the schedule that modifies the design infiltration volume flow rate

T_{zone}, T_{odb} = the indoor and outdoor air dry-bulb temperatures, °C

The EnergyPlus model of the High-Rise Apartment prototype building uses I_{design} = 0.57 L/(s•m^2) (0.11 CFM/ft^2), $F_{schedule}$ = always 1, coefficients *A, B,* and *D* = 0, and *C* = 0.224. This simplified approach to modeling infiltration in the prototype building models does not consider the effects of indoor-outdoor temperature differences, HVAC operation, or wind direction on air leakage. In contrast, the online calculator uses CONTAM to estimate air leakage rates, which considers all these factors mentioned while accounting for multizone building airflow physics.

Table 3 lists the four levels of air leakage rates that were used in the simulations of the High-Rise Apartment building. These calculations assume the air leakage is equally distributed over all exterior surfaces and include the slab and below-grade envelope area in the normalization of the air leakage rate, which is why they are referred to as 6-sided leakage values. The 6-sided value is used as the requirement in many building codes and standards; however, the CONTAM and EnergyPlus models assume no air leakage through any part of the exterior envelope that is not exposed to ambient air. The baseline value in Table 3 was calculated using the average building envelope airtightness for commercial buildings reported by Emmerich et al. 2005, 9 L/(s•m^2) (1.8 CFM/ft^2) at 75 Pa (0.3-inch water column) for a 5-sided envelope. The baseline of 7.7 L/(s•m^2) (1.5 CFM/ft^2) at 75 Pa (0.3-inch water column) was obtained by multiplying the 5-sided value by the 5-sided-to-6-sided envelope area ratio of the High-Rise Apartment building prototype. Table 3 also lists three target levels for improved airtightness at 75 Pa (0.3-inch water column): 2 L/(s•m^2) (0.4 CFM/ft^2) is the most stringent of three options and is found in the 2015 International Energy Conservation Code (IECC 2015) because it involves a blower door test, whereas the other two options are based on laboratory tests using ASTM E2357 and ASTM E2178. The airtightness required by the US Army Corps of Engineers is 1.25 L/(s•m^2) (0.25 CFM/ft^2) (USACE 2012); and 0.25 L/(s•m^2) (0.05 CFM/ft^2) is used to estimate performance at lower leakage rates. Emmerich and Persily (2014) analyzed the NIST US commercial building air leakage database and found that the 79 buildings categorized as having air barriers had an average 6-sided leakage of 1.39 L/(s•m^2) (0.27 CFM/ft^2) at 75 Pa (0.3-inch water column), which was 70% below the average leakage rate of the 290 buildings without air barriers (i.e., 4.33 L/(s•m^2) (0.85 CFM/ft^2) at 75 Pa (0.3-inch water column)); the former rate is similar to the second target level above. Zhivov (2013) reported the average 6-sided leakage for a set of 285 new and retrofitted military buildings constructed to the USACE specifications to be 0.9 L/(s•m^2) (0.18 CFM/ft^2).

Table 3. Assumed 6-Sided Building Envelope Airtightness Levels for the High-Rise Apartment

Case	Air Leakage Rate at 75 Pa (0.3-inch water column), L/(s•m^2) (CFM/ft^2)	Source
Baseline	7.7 (1.5)	Emmerich et al (2005)
1	2.0 (0.39)	IECC (2015)
2	1.25 (0.25)	USACE (2012)
3	0.25 (0.05)	

The annual total amount of moisture that is transported through the building envelope due to air leakage (M_W) is calculated using Eq. (2).

$$M_W = \sum_{i=1}^{n} \sum_{h=1}^{8760} \dot{m}_{a_{i,h}} W_h \,, \tag{2}$$

where

$\dot{m}_{a_{i,h}}$ = hourly mass flow rate due to air leakage for each zone

W_h = hourly humidity ratio of the outdoor air

i = zone number

h = hour of the year

n = number of zones (e.g., the High-Rise Apartment building model has 90 zones)

Figure 5 shows the annual moisture transfer through the building envelope due to air leakage at four envelope airtightness levels for four cities (Miami, FL; Chicago, IL; Phoenix, AZ; and Winnipeg, Canada) for the High-Rise Apartment building. These cities cover the range in annual moisture transfer for all cities included in the calculator. Figure 5 also shows the quadratic regression equations for each city and the coefficients of determination for the regression equation. Similar equations were derived for each city and used to calculate the moisture transfer as a function of building envelope airtightness for each building type. For the High-Rise Apartment building, the annual moisture transfer at a building envelope leakage rate of 7.7 L/(s•m^2) (1.5 CFM/ft^2) is 546, 367, 288, and 175 metric tons (1.2e+6, 8.1e+5, 6.3e+5, and 3.9e+5 lbs.) (118, 79, 62, and 38 kg/m^2 (24.2, 16.2, 12.7, and 7.8 lbs./ft^2) of exterior envelope area) for Miami, Chicago, Winnipeg, and Albuquerque, respectively.

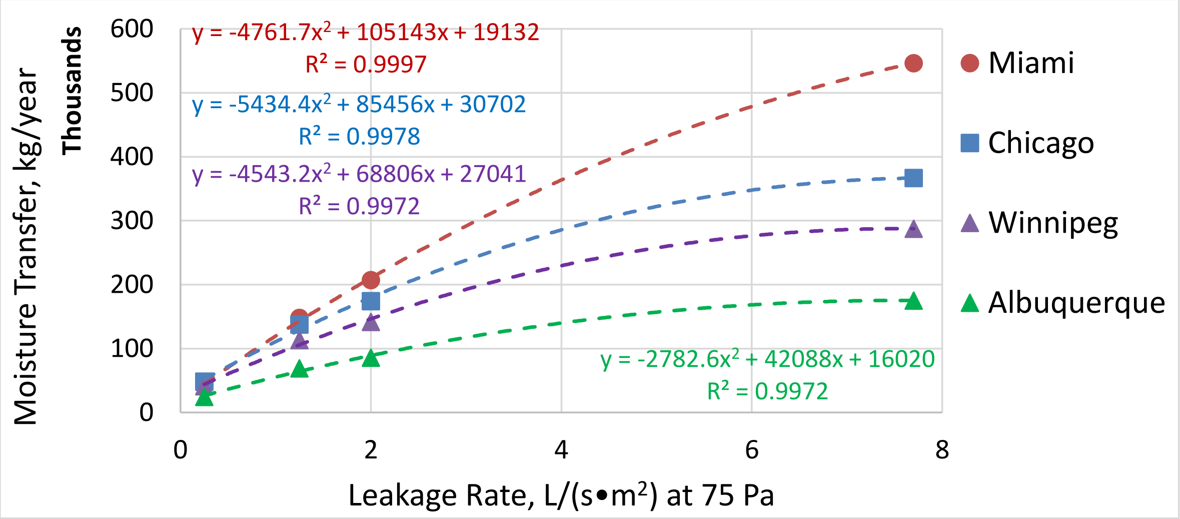


Figure 5. Annual moisture transfer through the building envelope due to air leakage at various envelope airtightness and locations. Note: 1 kg = 2.2 lb., 1 L/(s•m^2) = 0.2 CFM/ft^2, 75 Pa = 0.3-inch water column.

SUMMARY

In 2016–2017, ORNL, NIST, the ABAA, and the CERC BEE collaborated to develop an online calculator that uses the simulation results of the whole building energy simulation tool EnergyPlus and the multizone airflow simulation tool CONTAM. In 2018–2019, the calculator was expanded to add moisture transfer calculations because air leakage through the building envelope can have a significant impact on the amount of moisture transfer. Four more commercial building types were also added to the existing database of three building types. This paper describes the procedure used to calculate moisture transfer due to air leakage. This paper supplements Shrestha et al. 2016, which describes the calculation of energy savings due to the increase in envelope airtightness.

The procedure used in the online energy savings and moisture transfer calculator is different from other methods commonly used in energy analysis in that it uses hourly air leakage rates that are estimated by considering key variables such as building leakage rate, weather conditions, and HVAC operation. The calculator provides energy and costs savings and reduction in moisture transfer as a function of building envelope airtightness for the DOE commercial prototype buildings in 52 cities in the United States, five cities in Canada, and five cities in China. To demonstrate the moisture transfer calculations, the paper presents an example of how annual moisture transfer at an envelope leakage rate of 7.7 L/(s•m^2) (1.5 CFM/ft^2) at 75 Pa (0.3-inch water column) could be reduced by between 47% and 90% if the envelope leakage rate were reduced to 1.25 L/(s•m^2) (0.25 CFM/ft^2) at 75 Pa (0.3-inch water column) in the seven building types in Chicago.

The calculator is a powerful, credible, and easy-to-use tool that designers and contractors can utilize to estimate the benefits of reducing air leakage. These benefits include energy and cost savings in addition to a reduction in moisture transfer through the building envelope, as well as indoor air quality benefits not analyzed in this paper.

ACKNOWLEDGMENTS

The authors would like to thank the US Department of Energy and the ABAA for funding this research. This manuscript has been co-authored by UT-Battelle, LLC, under Contract No. DE-AC05-00OR22725 with the US Department of Energy.

REFERENCES

Armstrong, M., W. Maref, M.Z. Rousseau, W. Lei, and Nicholls, M. 2010. Effect of the air and vapour permeance of exterior insulation on the flow of moisture in wood frame walls in a cold climate. International Conference on Building Envelope Systems and Technologies. Vancouver, Canada, 2010

ASHRAE 90.1-2013. Energy Standard for Buildings except Low Rise Residential Buildings; ANSI/ASHRAE/IES Standard 90.1-2013; ASHRAE: Atlanta, GA, USA, 2013.

ASHRAE Handbook—Fundamentals. 2017. ASHRAE: Atlanta, GA, USA

ASTM E2357-11, Standard test method for determining air leakage of air barrier assemblies. West Conshohocken, PA: ASTM International.

ASTM E2178-13, Standard test method for air permeance of building materials. West Conshohocken, PA: ASTM International.

Belleudy, C., A. Kayello, M. Woloszyn, and H. Ge. 2015, Experimental and numerical investigations of the effects of air leakage on temperature and moisture fields in porous insulation. Building and Environment 94 (2015) 457-466

Crawley, D.B., J.W. Hand, M. Kummert, and B.T. Griffith.2008. Contrasting the capabilities of building energy performance simulation programs. *Building and Environment* 43(4):661–73

Deru, M.; Field, K.; Studer, D.; Benne, K.; Griffith, B.; Torcellini, P.; Halverson, M.; Winiarski, D.; Liu, B.; Rosenberg, M. 2011. DOE Commercial Reference Building Models for Energy Simulation–Technical Report; National Renewable Energy Laboratory: Golden, CO, USA

DOE. 2014. Windows and Building Envelope Research and Development: Roadmap for Emerging Technologies. Washington DC: US Department of Energy. http://energy.gov/sites/prod/files/2014/02/f8/BTO_windows_and_envelope_report_3.pdf

DOE. 2019. Commercial Prototype Building Models. Retrieved May 2019 from https://www.energycodes.gov/commercial-prototype-building-models.

Dols, W.S., Polidoro, B. 2016. CONTAM User Guide and Program Documentation; Technical Note 1887; National Institute of Standards and Technology: Gaithersburg, MD, USA

Emmerich, S., Persily, A., and McDowell, T.P. 2005. Impact of commercial building infiltration on heating and cooling loads in U.S. office buildings. Presented at 26th AIVC Conference, Brussels, September 21–23

Emmerich, S.; Persily, A. 2014. Analysis of US commercial building envelope air leakage database to support sustainable building design. Int. J. Vent. 2014, 12, 331–344

Goel, S., R. Athalye, W. Wang, J. Zhang, M. Rosenberg, Y. Xie, R. Hart, and V. Mendon. 2014. Enhancements to ASHRAE Standard 90.1 prototype building models. PNNL-23269. Richland, WA: Pacific Northwest National Laboratory

Gowri, K., D. Winiarski, and R. Jarnagin. 2009. Infiltration modeling guidelines for commercial building energy analysis. PNNL-18898. Richland, WA: Pacific Northwest National Laboratory

IECC. 2015. International Energy Conservation Code. https://codes.iccsafe.org/content/IECC2015?site_type=public

Janssen, A. and Hens, H. 2003. Interstitial condensation due to air leakage: A sensitivity analysis. Journal of Building Physics, Vol. 27 (1) (2003) 15-29

Ng, L.C., A. Musser, S.J. Emmerich, and A.K. Persily. 2012. Airflow and indoor air quality models of DOE reference commercial buildings. Technical Note 1734. Gaithersburg, MD: National Institute of Standards and Technology

Ng, L.C.; Ojeda Quiles, N.; Dols, W.S.; Emmerich, S.J. 2018. Weather correlations to calculate infiltration rates for U.S. commercial building energy models. Build. Environ. 2005, 127, 47–57

Polidoro, B., L.C. Ng, and W.S. Dols. 2016. CONTAM results export tool. Technical Note 1912. Gaithersburg, MD: National Institute of Standards and Technology

Shrestha, S.; Hun, D.; Ng, L.; Desjarlais, A.; Emmerich, S.; Dalgleish, L. 2016. Online airtightness savings calculator for commercial buildings in the United States, Canada, and China. In Proceedings of the Thermal Performance of the Exterior Envelopes of Whole Buildings—13th International Conference, Clearwater, FL, USA, 4–8 December 2016

Tenwolde, A. and W.B. Rose. 1996. Moisture control strategies for the building envelope. Journal of Building Physics, Vol. 19 (3) (1996) 206-214

USACE. 2012. Air Leakage Test Protocol for Building Envelopes. US Army Corps of Engineers

Zhivov, A., Herron, D., Durston, J.L., Heron, M., and Lea G. 2013. Air Tightness in New and Retrofitted US Army Buildings. AIVC Workshop

Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

INFRASTRUCTURE FOR MODEL BASED ANALYTICS FOR MANUFACTURING

Sanjay Jain

George Washington University
Department of Decision Sciences
Funger Hall #415, 2201 G Street NW
Washington, DC 20052, USA

Anantha Narayanan

University of Maryland
Department of Mechanical Engineering
Glenn L. Martin Hall, 4298 Campus Dr.
College Park, MD 20742, USA

Yung-Tsun Tina Lee

National Institute of Standards and Technology
Engineering Laboratory
100 Bureau Drive
Gaithersburg, MD 20899, USA

ABSTRACT

Multi-resolution simulation models of manufacturing system, such as the virtual factory, coupled with analytics offer exciting opportunities to manufacturers to exploit the increasing availability of data from their corresponding real factory at different hierarchical levels. A virtual factory model can be maintained as a live representation of the real factory and used to highly accelerate learning from data using analytics applications. These applications may range from machine level to manufacturing operations management level. While large corporations are already embarking on model based analytics initiatives, small and medium enterprises (SMEs) may find it challenging to set up a virtual factory model and analytics applications due to barriers of expertise and investments in hardware and software. This paper proposes a shared infrastructure for virtual factory model based analytics that can be employed by SMEs. A demonstration prototype of the proposed shared infrastructure is presented.

1 INTRODUCTION

Multiple trends are coming together to create exciting opportunities across all aspects of human lives. These trends include increasing computing power with accompanying reduction in its cost, increasing connectivity and speed across internet, increasing capabilities for data collection through sensors, and increasing access to data and applications via cloud and fog computing technologies. These trends have largely relaxed the constraints of computing power and data availability and enabled rapid growth in application and further development of technologies such as Internet of Things (IoT), artificial intelligence (AI) and analytics. The application of simulations and mathematical optimization techniques are rapidly growing too, taking advantage of the same trends. A number of these technologies have been around for years but were constrained in their applications due to lack of infrastructure of computing power, data access, and connectivity. For example, artificial intelligence attracted a lot of attention in 1980s building on special purpose languages such as PROLOG and LISP, special purpose hardware such as Texas Instruments Explorer workstation, and deployment of expert system applications for such varied fields as medical advice (Shortliffe 1986) and manufacturing scheduling (Jain et al. 1989). However, such developments plateaued with their growth constrained by the infrastructure limitations.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun. "Infrastructure for Model Based Analytics for Manufacturing." Paper presented at 2019 Winter Simulation Conference, National Harbor, MD, US. December 08, 2019 - December 11, 2019.

The ongoing rapid growth in infrastructure has not only provided an opportunity for development of earlier developed technologies, it has also spurred further developments across most of them. AI in particular has gained from a shift from mimicking human intelligence and rule based approaches of 1980s to now building on machine learning based analytics. Analytics applications employ increasingly rigorous and advanced applications involving large amount of computations. For example, Neural Networks are getting stacked for deep learning applications. Mathematical optimization and simulation applications have seen new developments based on the advancements in infrastructure. Simulations software today allow quick executions of large models that integrate multiple paradigms such as system dynamics and discrete event representations, compared to a couple of decades ago where applications of even single paradigm models were limited in size due to execution speed limitations.

There are multitude of efforts for developing and deploying standalone applications based on the four technologies listed above, namely AI, analytics, mathematical optimization, and simulation. Efforts are also beginning to be reported that synergistically employ these technologies together. This paper proposes an infrastructure for an application that brings together simulation and analytics technologies for supporting manufacturing operations management and machine level decisions. Majority of recent and under development analytics applications analyze real data to identify patterns and develop insights to support decision making. However, the knowledgebase of such applications is restricted to analyzing scenarios that occur in real life and thus their predictions are credible within that envelope only. Simulation applications on the other hand can generate credible outputs for a range of what if scenarios. Using the simulation as a data-generator for a range of scenarios and using that data to train analytics application can increase the prediction envelope of the analytics application by several fold. Such use of simulation models to generate data, which is then analyzed by data driven analytics applications, has been referred to as model based analytics.

This paper reports on the next step in bringing model based analytics closer to actual implementation. The progression of the work has been reported in successive years at this conference. Jain et al. (2017) presented initial work for model based analytics by linking a virtual factory prototype to a Neural Network (NN) for developing a meta model for manufacturing order promising. Jain et al. (2018) took the model based analytics further by linking virtual factory prototype to a Gaussian Process Regression (GPR) application and using the capability to compare NN and GPR for the order promising function. This paper proposes infrastructure for deploying the model based analytics capability to facilitate its eventual application by small and medium enterprises (SMEs). The emphasis is on SMEs since they generally lack the resources to develop and implement advanced technologies.

The next section briefly reviews recent literature for similar efforts. Section 3 presents a modeling and analysis framework that has been developed with a focus on machine learning based analytics applications for manufacturing. The framework is enhanced in Section 4 to include simulation models and to link them with analytics applications to enable (simulation) model based analytics. A prototype implementation of the proposed infrastructure is described in Section 5. Section 6 concludes the paper with discussion of future directions.

2 RELATED WORK

2.1 Virtual Factory

This paper uses the definition of a virtual factory as a multi-resolution simulation model of a manufacturing system capable of supporting analysis at different levels of hierarchy with interfaces to real factory and analytics applications. The definition largely overlaps with those of digital twin of factory, shop floor digital twin, and digital factory used by some authors (for example, see Garetti et al. 2012). Reported work over the last 2 years relevant to this definition of virtual factory are briefly reviewed in this section. The work focusing on virtual reality aspects or collaboration across manufacturers to make a “virtual factory” is not included here. Readers are referred to Jain et al. (2017) for relevant literature prior to 2017.

A number of efforts report development and implementation of virtual factory models with increasingly diverse applications and identify challenges. Hwang et al. (2017) use a virtual factory model to demonstrate and validate the use of an IoT based performance measurement system. Brenner and Hummel (2017) describe a prototype implementation of a digital twin in a learning factory setting. Modoni et al. (2018) present the digital twin of a factory as a digital factory and also as a virtual factory and highlight its benefits and technical challenges including the lack of interoperability of supporting software systems. Caggiano and Teti (2018) present a digital factory model that allows machine level modeling using a 3D simulation and cell level flow using a discrete event simulation. These efforts indicate continued development of the virtual factory concept with varied applications.

2.2 Model Based Analytics for Manufacturing

There appears to be little work reported that employs model based analytics for manufacturing applications. Giri et al. (2012) propose use of model based analytics for management of power grid and point to its capability for what-if analysis and coming up with corrective actions as a major benefit over measurement based analytics. Kajmakovic et al. (2018) propose to use model based analytics for predictive fail safe systems in industrial operations. Some authors identify the approach as simulation based analytics. Biller et al. (2017) modified simulation models of Silicon Carbide manufacturing operations and analyzed the outputs to address the challenge of limited data from a real facility. Ji and AbouRizk (2018) utilize simulation based analytics for decision support for pipe welding quality management in industrial construction. The opportunities offered by combination of simulation and data driven analytics models are being recognized and its increasing use is anticipated.

2.3 Model Based Analytics Infrastructure

The infrastructure for model based analytics and virtual factory has recently drawn attention of researchers. Chen and Lin (2017) support use of Factory Simulation as a Cloud Service (FSaaCS) for SMEs and identify two major issues, the need to convert models for different simulation systems available on different clouds and estimating the simulation time. They focus on load balancing approaches for FSaaCS. Coronado et al. (2018) describe a Manufacturing Execution System (MES) based on MTConnect protocol and cloud based technologies suitable for SMEs that can be used to develop a shop floor digital twin, an overlapping concept to virtual factory discussed above. The proposed MES could be one building block that SMEs can employ for developing the model. Further development can bring in analytical tools for model based analytics.

A few efforts have been reported outside the manufacturing domain. Lee et al. (2013) present a model based analytics service available via cloud for building energy consumption. He et al. (2018) describe a multi-tier fog computing structure for model based analytics of large scale IoT for smart cities. Gausemeier et al. (2011) discuss integration of manufacturing with control engineering and information services and describe a general procedure model for integrative development of mechatronic products. Lee et al. (2014) propose a framework for self-aware and self-maintained machines for Industry 4.0, which includes cyber physical systems and decision support systems, and address the trends of manufacturing service transformation in big data environment. To our knowledge, there is no framework that integrates simulation into the design and analytics aspects, thus expanding the data driven capabilities beyond the status quo.

Overall, this brief review indicates that there is recognition that the advancements in simulation and data analytics are offering exciting potential for their combined application as model based analytics. Integration frameworks need to be developed or extended to incorporate such combinations. There is also recognition of an emerging need to facilitate the application of model based analytics through development of infrastructure particularly for SMEs.

3 FRAMEWORK FOR DATA DRIVEN MODELS

Data analytics (DA) allows identifying performance improvements across multiple levels of manufacturing system functionality. DA techniques have been applied in manufacturing for many years (Harding et al.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun. "Infrastructure for Model Based Analytics for Manufacturing." Paper presented at 2019 Winter Simulation Conference, National Harbor, MD, US. December 08, 2019 - December 11, 2019.

2006). Most of these DA applications, however, are addressed to very specific issues under specific conditions (Sharp et al. 2018). Furthermore, DA applications are prohibitively complex and expensive, especially to SMEs. This is because DA techniques often require expertise in data collection, data analysis, machine learning, and decision optimization. A typical SME cannot afford to have a DA expert on staff. Thus, manufacturers need an enhanced decision-support facility so that analysis results at the process level can be elevated and used to influence enterprise-level decision making faster and better. The manufacturing industry can benefit greatly if such facility can 1) represent a wide range of manufacturing problems, 2) connect them to appropriate DA solutions, and 3) translate DA results into decisions that impact manufacturing operations across different levels of hierarchy. To address the need, a model-based analytics framework for manufacturing has been developed (Narayanan and Lee 2018) and is shown in Figure 1.

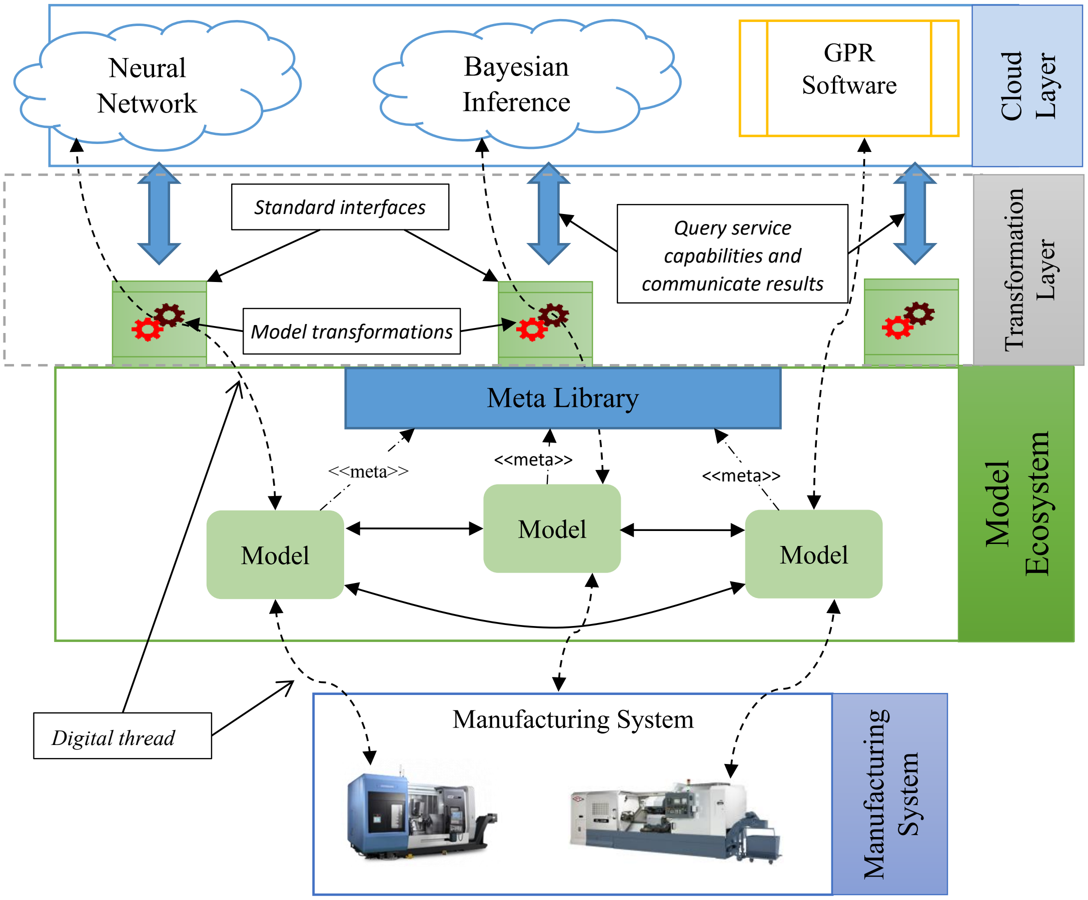


Figure 1: Model-based analytics framework for manufacturing.

The framework consists of a set of software to connect independent cloud and third-party DA applications or services to core manufacturing models. The framework supports both model integration and service integration across the entire hierarchy. The goal of the framework is to help the manufacturing industry to match their requirements to appropriate analysis services. The framework consists of four major layers described below.

- The manufacturing system layer, representing the physical system, includes the physical factory, physical sensors, and other data generators. The physical system may be composed of multiple subsystems organized as different lines, departments, cells, etc. At lower levels of the hierarchy, various machine tools and equipment are used to performs steps in the process plan for different products.
- The model ecosystem layer includes two major components: the domain-specific modeling environment and the library of meta-models. This layer provides the technical foundations for connecting a variety of analysis tools and services. The modeling environment will simplify system specification for manufacturing operations across different levels of hierarchy. It takes the domain-specific modeling approach to model abstracts in digital representations for manufacturing systems. The digital, manufacturing-domain models are then converted into analytical-domain models using model transformation to facilitate DA applications. Meta models are used to describe various aspects of manufacturing systems, based on the essential elements and rules of the domain of the individual physical systems; they are presented in a standardized way to facilitate model exchange and integration (OMG 2016).
- The transformation layer includes a set of model transformations or software tools that transfer system models into analysis models. It utilizes mapping algorithms that identify specific entities in the manufacturing-system domain and produce a corresponding entity in the analysis domain. The analysis model will be used to provide a solution to the manufacturing problem. When possible, the analysis model can be generated in a standard format, such as the Predictive Model Markup Language (PMML) (Guazzelli 2019), which can then be used with a variety of off-the-shelf analysis tools.
- The cloud layer includes various third-party services addressing various analytics needs. The third-party services can be standalone or cloud-based applications. The model ecosystem is connected to the cloud layer to allow various third-party tools to be used to perform analyses based on the system models.

To support the interaction among the layers, the framework provides additional components such as model library, standard interfaces, digital thread, and analysis discovery services. The model library includes pre-built system models to accelerate system specification, analysis models for various analysis objectives, reference data sets for specific scenarios, and possibly, extensions based on custom scenarios and data by individual organizations. Standard interfaces, based on standardized protocols and guidelines, are provided to make it easier for manufacturers to use available analytics services on their system models. The framework uses digital thread (Society of Manufacturing Engineers 2011) mechanism to connect digital representations and their corresponding physical entities. The digital thread allows manufacturers to trace analysis results back to actual physical components of the system. This traceability is necessary and plays a key role in decision making. The analytics discovery service guides manufacturers to the appropriate analysis services based on their needs. The discovery service is a very important component of the framework since the manufacturing users often are not experts in analytics. The service discovery interface will make it easy for manufacturers to find implementations that will provide the analytics service appropriate for solving the relevant analysis problem. The implementation of such service will require a high level of expertise in both the manufacturing and the analytics domains.

The core parts of the framework are the modeling techniques and the communication interface. Advanced modeling techniques are used to define intuitive and robust abstractions and interfaces for system specification and problem formulation. The framework is being developed using a service-oriented approach. The core modeling abstractions are being developed in a standardized way, e.g., the Core Manufacturing Simulation Model standard (Lee et al. 2011) has been used to facilitate the specification of complex manufacturing systems. Standard interfaces, e.g., PMML, are provided to communicate with third-party services for various analysis tasks. The analysis services are being built relying on the accurate transfer of relevant information through the standard interfaces. In the initial implementation, the framework is envisaged to carry only the basic items in each component. It is expected more component items will be developed through a collaborative effort by the community of researchers and practitioners.

4 PROPOSED INFRASTRUCTURE

In this section, we propose enhancements to the framework described in Section 3 to include simulation models in the framework and have their results be analyzed by data driven models.

4.1 Enhanced Framework for Model Based Analytics

Section 3 described the framework for data driven models for advanced manufacturing. Two of the essential components of the framework are the model ecosystem, which holds the digital representations of the manufacturing system, and the transformation layer, which enables the creation of data driven models from the digital representations. We extend these components to support simulation models. We extend the transformation layer to be able to automatically generate simulation models from the system representations in the model ecosystem. The implementation details are described in Section 5. Figure 2 shows the extended framework to support simulation models.

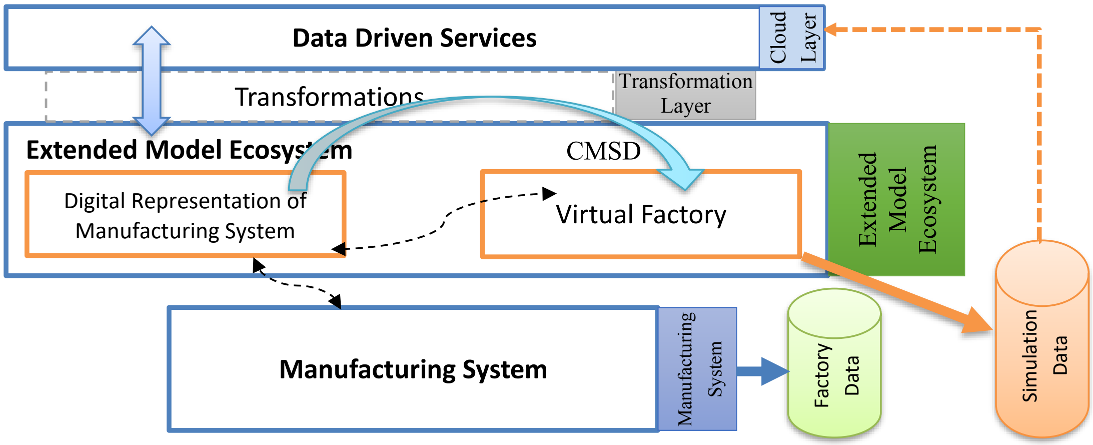


Figure 2: Extended framework for simulation model based analytics.

The goal of the extended framework is to provide the ability to easily configure simulations of a manufacturing system. The extended framework will allow users to make arbitrary changes to the digital representation of the system to easily create new configurations of the system. A model transformation component will automatically generate a simulation model from the digital representation in a standard CMSD (Riddick and Lee 2010) format. This allows users to rapidly generate multiple simulation models for various configurations. The standard simulation model may then be imported into a variety of simulation tools to run simulations. These simulations can be used for various purposes, such as generating synthetic data for new data driven models, or for performing other types of optimizations. In the subsections below, we discuss these applications in more detail.

4.2 Virtual Factory Model

The “model” employed for model based analytics is a virtual factory in the context of manufacturing. The virtual factory is envisaged as a multi-resolution model that allows representation and analysis of the corresponding real factory at different levels of hierarchy. In a real factory, the performance can be analyzed at various levels such as machines, cells, departments, and the entire factory including their interactions with supporting system using the data streams from various sensors and data collection systems. The virtual factory would allow analysis at the desired level of resolution similar to the real factory except it would have the advantage to create and analyze a range of future potential scenarios. The virtual factory will need to be closely linked to the real factory to be its high fidelity representation. The virtual factory will also need to be interfaced with data analytics applications similar to the real factory for this purpose. The virtual factory is proposed to reside in the extended model ecosystem to gain from defined interfaces to the real manufacturing system and the data driven models in the framework.

The virtual factory will have a synergistic relationship with data driven applications. The primary intent is of course to have the virtual factory generate data for model based analysis. However, the analytics applications will help the virtual factory in return by improving the data used for simulations. For example, the data for machine failures can be continuously tracked and periodically analyzed to improve the distributions used to represent them in the virtual factory.

4.3 Support for Model Development

The framework provides an intuitive domain specific modeling environment that makes it easy for practitioners to develop digital models that closely and accurately represent their manufacturing systems. The modeling environment enforces rules that ensure that the models are sound, and allows for error checking. Digital threads trace relationships between model elements and the physical component that they represent. This allows us to develop rule based transformations that can automatically create other types of models from these models. One example of such a model transformation allowed us to generate neural networks for predicting energy consumption from milling machine models (Lechevalier et al. 2014).

In the extended framework, we implemented such a model transformation to generate a CMSD file as a standard representation of a simulation model. This transformation is completely automated, and allows users to make changes in the factory model and quickly generate simulation models. The CMSD file may be imported into a simulation tool such as AnyLogic to execute the simulation. More details of the implementation are presented in Section 5.

4.4 Support for Analytics

As stated before, the transformation layer of the framework implements various model transformations to generate analytic models from the system model. Data from the factory is used to train the analytic models. However, in many cases data from the factory is not available. For example, if the factory is newly set up, there may not be sufficient historic data. Also, the user may want to experiment with various factory configurations, and data may not be available for all of these new configurations. In such cases, it is beneficial to have this data generated synthetically through simulations. Figure 2 shows how simulations can be executed to generate simulated data, which can then be used to train the analytic models.

5 INFRASTRUCTURE PROTOTYPE

In this section, we present our preliminary implementation of the framework, and describe its use through an example.

5.1 Implemented User Modeling Environment

The model ecosystem shown in Figures 1 and 2 was implemented using the Generic Modeling Environment (GME) (Ledezci et. al. 2001). GME is an open source tool for creating configurable modeling environments. We developed a meta-model (abstract model) in GME that describes the main concepts and modeling rules of the model ecosystem. The meta-model defines the rules based on which models can be built. The concepts and relationships in the meta-model are "instantiated" to build models. The models represent actual items in the physical systems in that domain. Figure 3 shows a screenshot of the domain specific modeling environment in GME. It shows an instantiated model of a manufacturing process, showing the process parameters, metrics and variables. The tree structure on the right of Figure 3 represents the hierarchical structure of the entire system model. The environment we implemented for this framework allows the user to model the factory layout, parts manufactured, the process plans, and the resources used.

We implemented a model transformation that takes the factory representation shown in Figure 3, and generates a CMSD file. The CMSD file is an XML file that represents all the elements of the factory model necessary to create a simulation of the factory. We implemented a plug-in for the AnyLogic simulation tool, to read the standard CMSD file and generate an AnyLogic simulation model. We executed the simulation in AnyLogic, and generated data from the simulation. The generated data was then used to train machine learning models for cycle time predictions for the purpose of order promising. We generated a Neural Network (NN) and a Gaussian Process Regression (GPR) model. These models are described in Section 5.3.

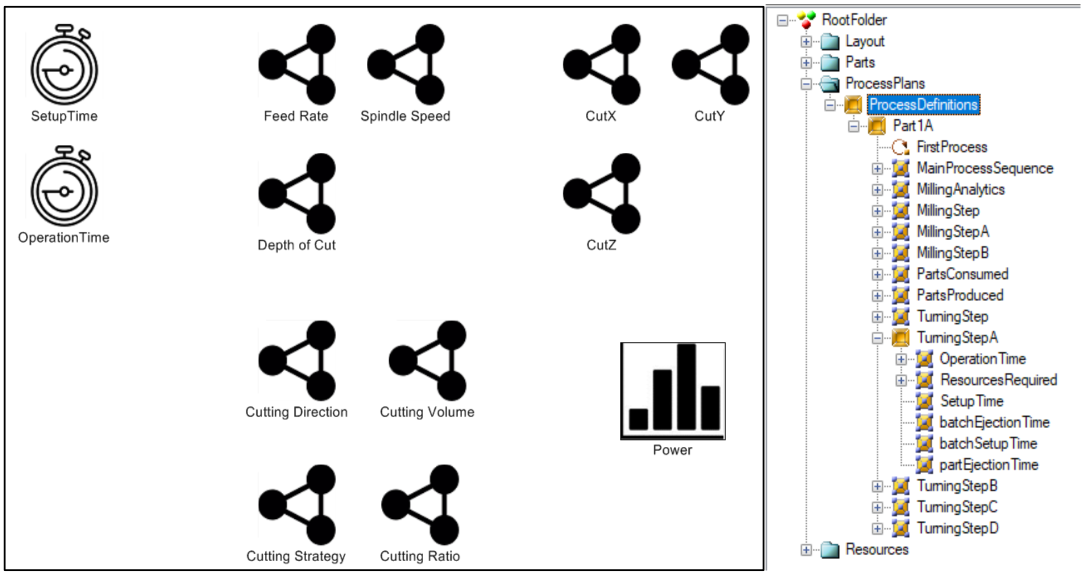


Figure 3: Model representation of a manufacturing process.

5.2 Implemented Virtual Factory Model

The virtual factory model is based on the scenario described in Jain et al. (2017) and Jain et al. (2018). The use case is based on order promising for a small job shop and hence the corresponding model may be referred to as a virtual job shop. The job shop produces parts from three different material types. The goal is to predict a shipment date at the time of the order, based on the material type chosen for the part, and the current load on the shop.

The job shop model was built in GME, in the domain specific modeling environment (DSME) described above. The DSME allows us to visually construct a hierarchical model of the job shop, and specify all the machine parameters within the model itself. The job shop model consists of a turning cell with four turning machines, and a milling cell with two milling machines. A process model is built in the DSME to describe the process flow. The model elements can be seen in the tree hierarchy on the right side of Figure 3. MillingStep and TurningStep specify the machine cells, and MainProcessSequence specifies the overall process sequence.

We implemented a translator tool that automatically generates a standard CMSD file from the shop floor model specified in the DSME. This CMSD file was then imported into the AnyLogic simulation tool (using another plugin that we implemented), to execute the simulation. The virtual factory prototype does have the capability to utilize multi-resolution models with the most detailed level capable of modeling machine dynamics. However, with the focus on analyzing job shop cycle times for this study, only the discrete event model of material flow across the shop was used. We generated synthetic data for the scenario from the simulation, which was then used to train a prediction model for order promising.

The advantage of building the model in the DSME and generating the simulation model, as opposed to building the simulation model directly in the simulation tool, is twofold. First, The DSME is designed to be used by a manufacturing domain expert, and uses concepts and visual representations that are easy to understand for a domain user. This can greatly accelerate model development, reduce model errors, and ease maintenance of models. Second, The DSME allows us to implement many different algorithms for various tasks in a more powerful and elegant way than does a simulation tool (which is restricted to simulation tasks). In particular, we can implement several model translators to generate other types of models from the specification in the DSME. For example, we implemented a Neural Network (NN) generator that can automatically generate a NN for a machine from its specification in the DSME (Lechevalier et al. 2014). This neural network can then be trained on data (real or synthetic) for that machine. Thus, from a single model in the DSME, we are able to generate and run a simulation, generate an NN model, and use the data from the simulation to train the NN. Figure 4 provides an overview of the infrastructure prototype.

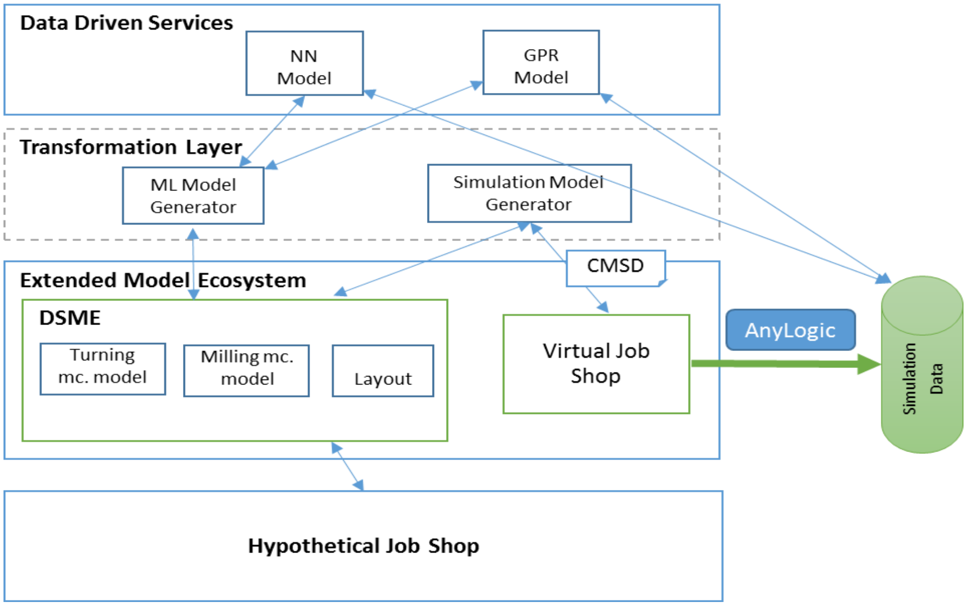


Figure 4. Overview of the implemented infrastructure prototype.

5.3 Implemented Analytics Applications

Based on the simulation model described above, we implemented two different machine learning models to predict throughput. We first trained a Neural Network (NN) model (Jain et al. 2017) for throughput prediction. We then generated a Gaussian Process Regression (GPR) model (Jain et al. 2018) for the same purpose, and compared it to the prediction accuracy of the NN model. We summarize these models here. For theoretical background on NNs and GPR, we refer the reader to Haykin (2004) and Rasmussen (2004).

The NN model takes a set of input variables and makes a prediction on a desired target variable. In our case, the target was “cycle time”, which was needed to estimate the expected time for order completion. The input variables included the material type, the number of parts in the order, and the current load on the system. The material type is one of aluminum, steel or titanium, and is represented using one-hot encoding, i.e., it is represented by three variables (one for each material), with a zero or one depending on the material chosen. The load on the system is modeled by a triplet (n_A, n_S, n_T), which denotes the parts of material type aluminum, steel, and titanium currently being processed in the system. For a new job shop, there is not sufficient real data to train a reliable model for predicting duration under many varying conditions. Our simulation model described previously overcomes this problem, by generating simulated data for a variety of shop conditions. The data generated by the simulation model is then used to train the NN model. The resulting NN model is capable of predicting an estimated duration for new orders, and can be used to give an estimated ship date when the shop is operational.

We also created a Gaussian Process Regression (GPR) model for duration prediction from the same simulated data. GPR is a probabilistic method of interpolation to determine the target value from the inputs. GPR produces a distribution of the expected target value, allowing us to account for uncertainty in the prediction. The GPR model uses the same inputs and outputs but provides a prediction window for the target based on the uncertainty. In input regions where sufficient training data was available to provide a reliable prediction, the uncertainty is small. Figure 5 shows a comparison of the NN and GPR models on simulated test data. The grey color bands around the GPR predictions represent the confidence bounds generated by GPR. More details of the comparison between these prediction models can be found in Jain et al. (2018).

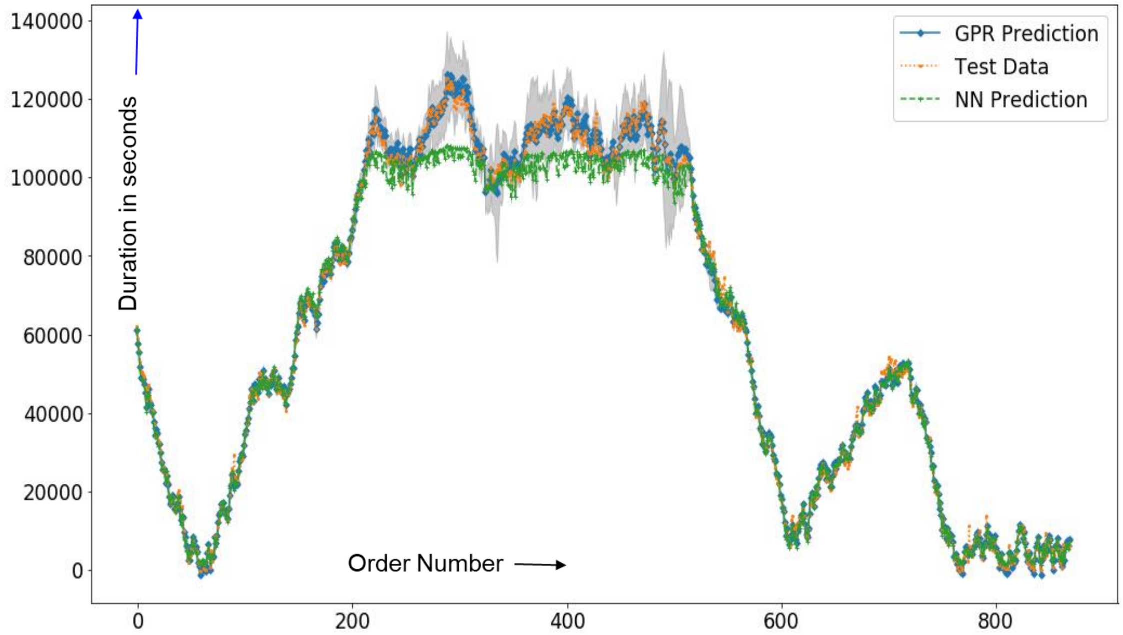


Figure 5: Comparing NN and GPR for cycle time predictions (Jain et al. 2018).

6 CONCLUSION

This paper proposed an infrastructure for model based analytics with emphasis on orienting the capability for SMEs. A prototype implementation for the infrastructure has been described that uses a small virtual factory to generate cycle time performance data that is then analyzed by NN and GPR applications to develop a predictive model for order promising. The next proposed step is to use data from a larger real manufacturing system to set up the model based analytics capability and exercise it for addressing issues of interest to the decision makers. Future work may focus on successive iterations of enhancements in all the components of the framework including the interactions with manufacturing personnel, increasingly automated generations of virtual factory models, interfaces of virtual factory with real factory data systems, and interfaces of virtual factory for guided analytics driven by decision makers.

DISCLAIMER

No approval or endorsement of any commercial product by the National Institute of Standards and Technology (NIST) is intended or implied. Certain commercial software systems are identified in this paper to facilitate understanding. Such identification does not imply that these software systems are necessarily the best available for the purpose.

ACKNOWLEDGMENTS

The work of Sanjay Jain and Anantha Narayanan on this effort was supported by NIST under Cooperative Agreement Nos. 70NANB18H010 and 70NANB18H246, respectively.

REFERENCES

Biller, B., O. Dulgeroglu, C. G. Corlu, M. Hartig, R. J. Olson, P. Sandvik, and G. Trant. 2017. “Semiconductor Manufacturing Simulation Design and Analysis with Limited Data”. In *2017 28th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC)*, May 15th-18th, Saratoga Springs, New York, 298-304.

Brenner, B. and V. Hummel. 2017. “Digital Twin as Enabler for an Innovative Digital Shopfloor Management System in the ESB Logistics Learning Factory at Reutlingen-University”. *Procedia Manufacturing* 9:198-205.

Caggiano, A. and R. Teti. 2018. “Digital Factory Technologies for Robotic Automation and Enhanced Manufacturing Cell Design”. *Cogent Engineering* 5(1):1426676.

Chen, T. and C. W. Lin. 2017. “Estimating the Simulation Workload for Factory Simulation as a Cloud Service”. *Journal of Intelligent Manufacturing* 28(5):1139-1157.

Garetti, M., P. Rosa, and S. Terzi. 2012. “Life Cycle Simulation for the Design of Product-Service Systems”. *Computers in Industry* 63(4):361-369.

Gausemeier, J., R. Dumitrescu, S. Kahl, and D. Nordsiek. 2011. “Integrative Development of Product and Production System for Mechatronic Products”. *Robotics and Computer-Integrated Manufacturing* 27(4):772-778.

Giri, J., M. Parashar, J. Trehern, and V. Madani. 2012. “The Situation Room: Control Center Analytics for Enhanced Situational Awareness”. *IEEE Power and Energy Magazine* 10(5):24-39.

Guazzelli, A. 2019. What is PMML? KDnuggets, https://www.kdnuggets.com/faq/pmml.html, accessed 25th April 2019.

Harding, J. A., M. Shahbaz, and A. Kusiak. 2006. “Data Mining in Manufacturing: A Review.” *Journal of Manufacturing Science and Engineering* 128(4): 969-976.

Haykin, S. 2004. *Neural Networks: A Comprehensive Foundation*. Prentice Hall.

He, J., J. Wei, K. Chen, Z. Tang, Y. Zhou, and Y. Zhang. 2018. “Multitier Fog Computing with Large-Scale IoT Data Analytics for Smart Cities”. *IEEE Internet of Things Journal* 5(2):677-686.

Hwang, G., J. Lee, J. Park, and T. W. Chang. 2017. “Developing Performance Measurement System for Internet of Things and Smart Factory Environment”. *International Journal of Production Research* 55(9):2590-2602.

Jain, S., K. Barber, and D. Osterfeld. 1989. “Expert System for Online Scheduling”. In *Proceedings of the 1989 Winter Simulation Conference*, edited by E.A. MacNair, P. Heidelberger, and K.J. Musselman, 930-935. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Jain, S., A. Narayanan, and Y. T. Tina Lee. 2018. “Comparison of Data Analytics Approaches using Simulation”. In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1084-1095. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Jain, S., G. Shao, and S.-J. Shin. 2017. “Manufacturing Data Analytics Using a Virtual Factory Representation”. *International Journal of Production Research* 55(18):5450-5464.

Jain, S., D. Lechevalier, and A. Narayanan. 2017. “Towards Smart Manufacturing with Virtual Factory and Data Analytics”. In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 3018-3029. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ji, W. and S. M. AbouRizk. 2018. “Simulation-based Analytics for Quality Control Decision Support: A Pipe Welding Case Study”. *Journal of Computing in Civil Engineering* 32(3):05018002.

Kajmakovic, A., R. Zupanc, S. Mayer, N. Kajtazovic, M. Höffernig, and H. Vogl. 2018. “Predictive Fail-Safe Improving the Safety of Industrial Environments through Model-based Analytics on Hidden Data Sources”. In *2018 IEEE 13th International Symposium on Industrial Embedded Systems (SIES)*, June 6th – 8th, 2018, Graz, Austria, 1-4.

Lechevalier, D., A. Narayanan, and S. Rachuri. 2014. “Towards a Domain-Specific Framework for Predictive Analytics in Manufacturing”. In *2014 IEEE International Conference on Big Data (Big Data),* Oct. 27th-30th, Washington, DC, USA, 987–995.

Lechevalier, D., S.-J. Shin, S. Rachuri, S. Foufou, Y. T. Lee, and A. Bouras. 2019. “Simulating a Virtual Machining Model in an Agent-based Model for Advanced Analytics”. *Journal of Intelligent Manufacturing* 30(4):1937-1955.

Ledeczi, A., A. Bakay, M. Maroti, P. Volgyesi, G. Nordstrom, J. Sprinkle, and G. Karsai. 2001. “Composing Domain-Specific Design Environments”. *Computer* 34(11):44–51.

Lee, J., H. A. Kao, and S. Yang. 2014. “Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment”. *Procedia CIRP 16*, 3-8.

Lee, Y. M., L. An, F. Liu, R. Horesh, R., Y. T. Chae, R. Zhang, E. Meliksetian, P. Chowdhary, P. Nevill, and J. L. Snowdon. 2013. “Building Energy Performance Analytics on Cloud as a Service”. *Service Science* 5(2):124-136.

Lee, Y. T., F. Riddick, and B. Johansson. 2011. “Core Manufacturing Simulation Data – A Manufacturing Simulation Integration Standard: Overview and Case Studies”. *International Journal of Computer Integrated Manufacturing* 24(8):689-709.

Modoni, G. E., E. G. Caldarola, M. Sacco, and W. Terkaj. 2019. “Synchronizing Physical and Digital Factory: Benefits and Technical Challenges.” *Procedia CIRP* 79:472-477.

Narayanan, A., and Y. T. Lee. 2018. *Model-Based Approach towards Integrating Manufacturing Design and Analysis.* NIST Advanced Manufacturing Series 300-5. https://doi.org/10.6028/NIST.AMS.300-5, accessed 25th April 2019.

OMG, 2016. *Meta Object Facility, Version 2.5.1*. https://www.omg.org/spec/MOF/About-MOF/, accessed 25th April 2019.

Rasmussen, C. E. 2004. "Gaussian Processes in Machine Learning”. In *Advanced Lectures on Machine Learning,* 63-71. Berlin, Heidelberg: Springer.

Riddick, F. H. and Y. T. Lee. 2010. “Core Manufacturing Simulation Data (CMSD): A Standard Representation for Manufacturing Simulation-Related Information”. In *Fall Simulation Interoperability Workshop (Fall SIW),* September 20th-24th, Orlando, FL, USA, Paper ID: 10F-SIW-034, 1-9.

Sharp, M., R. Ak, and T. Hedberg, Jr. 2018. “A Survey of the Advancing Use and Development of Machine Learning in Smart Manufacturing”. *Journal of Manufacturing Systems* 48(Part C): 170-179.

Shortliffe, E. H. 1986. “Medical Expert Systems—Knowledge Tools for Physicians”. *Western Journal of Medicine* 145(6):830-839.

Society of Manufacturing Engineers. 2011. “Connecting the Digital Threads”. *Manufacturing Engineering* 146(6):33-36.

AUTHOR BIOGRAPHIES

SANJAY JAIN is an Associate Industry Professor in the Department of Decision Sciences, School of Business at the George Washington University. Before moving to academia, he accumulated over a dozen years of industrial R&D and consulting experience working at Accenture in Reston, VA, USA, Singapore Institute of Manufacturing Technology, Singapore and General Motors North American Operations Technical Center in Warren, MI, USA. His research interests are in application of modeling and simulation of complex scenarios including smart manufacturing systems and project management. His email address is jain@email.gwu.edu.

ANANTHA NARAYANAN is a Research Associate at the University of Maryland. He completed his PhD at Vanderbilt University, Nashville, TN, USA in 2008. His research interests are in data analytics and model based systems engineering. He is currently working as a guest associate in NIST’s Systems Integration Division of the Engineering Laboratory. His email address is ananth222@gmail.com~~anantha@umd.edu~~.

YUNG-TSUN TINA LEE is a computer scientist in the Systems Integration Division of the Engineering Laboratory at NIST. Currently, she co-leads the Data-Driven Decision Support for Additive Manufacturing Project of NIST’s Manufacturing Science for Additive Manufacturing Program. She is a co-editor of the Simulation Interoperability Standards Organization (SISO) Standards of Core Manufacturing Simulation Data (CMSD). Her email address is yung-tsun.lee@nist.gov.

Jain, Sanjay; Narayanan, Anantha Narayanan; Lee, Yung-Tsun. ”Infrastructure for Model Based Analytics for Manufacturing.” Paper presented at 2019 Winter Simulation Conference, National Harbor, MD, US. December 08, 2019 - December 11, 2019.

2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)

Effect of Federal Incumbent Activity on CBRS Commercial Service

Michael R. Souryal and Thao T. Nguyen
Communications Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, Maryland, U.S.
{souryal, thao.t.nguyen}@nist.gov

***Abstract*—Federal Communications Commission rules for the 3.5 GHz citizens broadband radio service (CBRS) permit commercial systems to share 150 MHz of bandwidth with federal and other incumbents. To protect the federal incumbents, a spectrum coordinator—the spectrum access system (SAS)—uses a standardized algorithm to suspend some commercial transmissions when a nearby incumbent (e.g., a shipborne radar) becomes active. Using propagation models based on those employed by the SAS for interference management, this paper quantifies the impact of federal incumbent activity on commercial service, in terms of both the numbers of transmissions affected and their service area. These metrics are also used to examine the tradeoff between commercial coverage and immunity to incumbent activity as a function of commercial base station antenna height. We find that, in some cases, the reduction in coverage from lowering antenna heights is more than offset by the gain in immunity to incumbent activity.**

I. Introduction

Regulatory rules for the citizens broadband radio service (CBRS) in the U.S. [1] define three tiers of access to the radio frequency (RF) spectrum from 3550 MHz to 3700 MHz (3.5 GHz band). The top tier, or highest priority users, are the incumbents; they include certain military operations, existing fixed satellite service earth stations, and grandfathered wireless broadband licensees. The second tier consists of new entrants to the band that have a priority access license but must not cause harmful RF interference to the top tier (incumbent) users. The third tier is for general authorized access not requiring a license; they must not cause interference to and must accept interference from the top two tiers.

The rules require that second and third tier access to the band be dynamically managed by a spectrum access system (SAS). CBRS devices (CBSDs)—the tier 2 and tier 3 devices that access the band (i.e., base stations and access points)—must receive authorization from the SAS prior to transmitting. Such authorizations include the maximum RF power at which the CBSD can transmit and the specific channel in the CBRS band on which to transmit.

The federal incumbents include both ground-based and shipborne military radars which may or may not be active at a given time. When informed of federal incumbent activity in a certain geographic area, a SAS must dynamically manage the aggregate interference of CBSDs so as not to exceed a predefined protection threshold anywhere in the protected area. To accomplish this interference management, the SAS may need to temporarily suspend certain commercial transmissions on the affected channel.

This paper is a study of the impact of federal incumbent activity on commercial use of the CBRS band. Employing some of the same propagation models mandated by CBRS standards for use by the SAS in managing access to the band, we quantify the impact of incumbent activity in terms of not only the numbers of affected CBSDs as in previous studies (e.g., [2]), but also in terms of their service area. Such an analysis enables the examination of various tradeoffs, such as increasing commercial service coverage with higher power transmissions or higher antennas versus lowering the impact of federal incumbent activity on commercial service.

Sections II and III describe the propagation models and commercial deployment model, respectively, used in this study. Section IV presents a quantitative analysis of the impact of incumbent activity on commercial service and examines the tradeoff in service coverage of antenna height. Finally, section V summarizes the results and draws conclusions.

II. Propagation Models

CBRS standards [3] specify the propagation models that a SAS must use to perform its functions, such as calculating a CBSD's coverage area and managing the aggregate interference of lower tier users to higher tiers.

A. CBSD Coverage

As part of its protection of tier-two users, a SAS is required to calculate a contour around each tier-two CBSD at which the received signal strength from that CBSD is equal to -96 dB relative to 1 mW (dBm) in a 10 MHz channel. To perform this calculation, the SAS uses a propagation model that is a hybrid of the irregular terrain model (ITM) (also known as Longley-Rice model) [4] and the extended Hata (eHata) model described in [5]. The eHata model is an extension of Hata's empirical formulae in both distance and frequency ranges and includes a number of site-specific corrections to the median attenuation, the details of which are provided in [5]. The hybrid model selects the appropriate loss model based on the distance between the transmitter and receiver. For distances less than 1 km, either free space path loss or an interpolation between free space and eHata is used. For longer distances, the loss is generally determined as the larger

Souryal, Michael; Nguyen, Thao T. "Effect of Federal Incumbent Activity on CBRS Commercial Service." Paper presented at IEEE International Symposium on Dynamic Spectrum Access Networks, Newark, NJ, US. November 11, 2019 - November 14, 2019.

of the ITM and eHata losses. However, if the transmitter has an effective height above 200 m or if it is deployed in a rural area, the loss is equal to the ITM loss, regardless of distance. Furthermore, if the CBSD is located indoors, 15 dB is added to the loss to account for building attenuation. Details of the CBRS hybrid model are provided in [3, R2-SGN-04]. An open-source reference implementation of the hybrid model is available [6], as are the terrain and other data used by the model [7].

B. *Federal Incumbent Protection*

While CBRS standards specify the hybrid model described above for the protection of tier two and certain grandfathered incumbents, it currently only specifies the use of ITM for the protection of federal incumbents. However, to better account for clutter and examine the tradeoffs in antenna height, in this study we employ the hybrid model that was used for federal incumbent protection in the 3.5 GHz exclusion zone analysis [5]. This hybrid model differs from that specified in the CBRS standards for tier-two protection and is used solely for calculating the aggregate interference to federal incumbents.

For the purposes of this study, we adapted the hybrid model of [5] in two important respects. First, because the exclusion zone analysis only considered rural antenna heights up to 6 m, it included clutter loss on all paths from rural areas. However, in this study, antenna heights in rural areas can be as high as 100 m, well above most clutter. Hence, we apply the antenna height threshold of 18 m used in [5] for urban and suburban areas to rural areas, as well: below 18 m we apply the same random, uniformly distributed rural clutter loss as in [5], and above 18 m we apply no clutter loss. Second, for the ITM portion of the hybrid model, we use the parameters specified in the CBRS standards [3] for consistency.

To illustrate the difference between the hybrid model and the CBRS standard model based on ITM alone, we consider a simulated deployment of CBSDs near a protected federal incumbent in Pensacola, Florida. In this example, all CBSDs have an antenna height of no more than 18 m. Fig. 1 shows histograms of the hybrid median path loss (in red), the CBRS standard's median path loss (in blue), and the difference (dB) between the two (in green). As shown in the figure, the difference ranges from −5 dB to 60 dB. Negative values are observed because indoor losses in the hybrid model can be as little as 10 dB, while the indoor loss in the standard model is fixed at 15 dB.

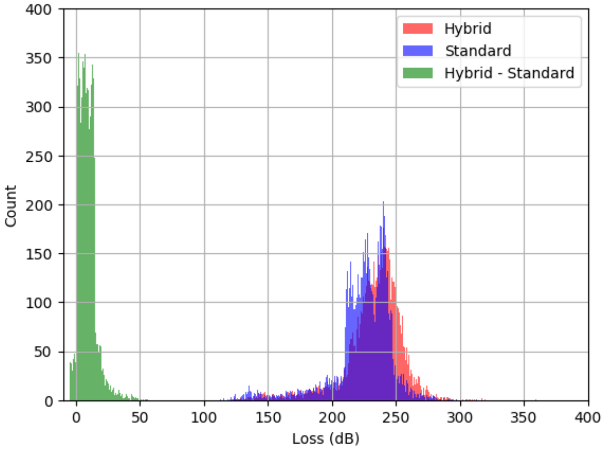


Fig. 1. Histograms of hybrid median path loss, CBRS standard median path loss, and their difference, for CBSDs near Pensacola

III. Commercial Deployment Model

The regulatory rules for the CBRS band specify two categories of CBSDs. Category A CBSDs are lower power devices with a maximum effective isotropic radiated power (EIRP) of 30 dBm/10 MHz and are typically installed indoors. Category B CBSDs are higher power devices (47 dBm/10 MHz maximum EIRP) and are professionally installed outdoors [1].

The simulated deployments used in this study are based on a model that was used by the National Telecommunications and Information Administration (NTIA) to size the coastal areas requiring federal incumbent protection, referred to as "dynamic protection areas" (DPAs) [8], as well as to determine the areas in which commercial service could potentially be affected, referred to in CBRS requirements as the "DPA neighborhood" [3].

The NTIA model, described in [2, Section III-A], generates the locations, antenna heights, and transmission powers of a simulated deployment of CBSDs around a given DPA. The numbers of CBSDs and their locations are a function of population and land classification. For this study, Category A CBSDs were placed as far as 250 km from the DPA boundary, and the higher-power Category B CBSDs were placed as far as 600 km from the boundary. Furthermore, all CBSD antennas were configured to be omnidirectional. Sample deployments generated with the NTIA model can be found at [6].

In addition, we adapted the deployment model to prevent overlap of coverage areas of each category of CBSDs. The rationale for this decision is that, since the deployment is for a given channel in the CBRS band—in other words, all transmissions of a deployment are co-channel—a SAS may try to limit the authorization of CBSD transmissions to avoid their mutual interference [9]. As such, the resulting deployment can be viewed as the effective deployment of CBSDs after SAS authorizations are made.

We apply a simple down-selection of CBSDs to obtain a non-overlapping set. Given a list of CBSDs generated by the NTIA deployment model, we step sequentially through each CBSD in the list and add it to the set if its −96 dBm/10 MHz contour does not intersect with the contour of any CBSD already in the non-overlapping set. We perform this selection separately for Category A CBSDs and for Category B CBSDs, resulting in two sets, each of which contains non-overlapping CBSDs. As such, the coverage of Category A CBSDs, which are considered to be indoors in the NTIA model, can overlap with that of the Category B CBSDs, which are outdoors. The rationale for independent selection of the two sets is that the Category A CBSDs can provide indoor coverage where Category B CBSDs may not.

We observe that our selection of non-overlapping contours does not necessarily maximize overall coverage and depends on the order of the starting list. Nevertheless, as an achievable solution, it provides a lower bound on the coverage of the optimal non-overlapping set.

Fig. 2 depicts the -96 dBm/10 MHz contours of both the original deployment and a non-overlapping deployment of CBSDs near the East 2 DPA (the red polygon in the figure) off the southeastern coast of the U.S. The total area of the union of the contours in the original deployment is $615\,390$ km^2, and the total area of the non-overlapping contours is $259\,623$ km^2.[1]

IV. Analysis

A. Deployment Scenarios

We evaluate the impact on CBRS commercial service of federal incumbent activity for three select DPAs: an offshore DPA off the southeastern coast of the U.S. (East 2), an offshore DPA near a naval base on the West Coast (West 14), and a point DPA on the U.S. Gulf Coast (Pensacola). We quantify the impact in terms of three metrics: (i) the number of authorized CBRS transmissions that are suspended when the DPA becomes active, (ii) the outer distance of the affected transmissions (a measure of the DPA neighborhood), and (iii) the total service area of the remaining active CBRS transmissions.

We compare these metrics under two deployment scenarios, one with unrestricted antenna heights and one with restricted antenna heights. In the latter scenario, we simply clip the height of each outdoor (Category B) CBSD in the original deployment to 18 m, i.e., $h_{\text{new}} = \min(h_{\text{orig}}, 18)$, where h_{orig} is the height of the CBSD in the original deployment in meters. The height restriction has the effect of shrinking the coverage area of a CBSD that has been clipped but, due in part to clutter loss at the lower antenna heights, reduces the likelihood that its transmission is suspended due to incumbent activity.

To illustrate the effect of the height restriction on coverage, Fig. 3 shows the histogram of the coverage area of the Category B CBSDs for each scenario of the East 2 deployment on a logarithmic scale. While the original deployment has CBSD coverage areas as high as 3500 km^2, the height restriction limits coverage areas to around 1000 km^2. In any case, the vast majority of CBSDs in the original deployment have a coverage area below 1000 km^2.

Nevertheless, as a result of the smaller coverage contours of some outdoor CBSDs, the selection of non-overlapping contours from the height-limited deployment yields a somewhat larger number of Category B CBSDs in the non-overlapping set. For the DPAs considered in this analysis, the increase ranges from $2\,\%$ to $16\,\%$ depending on the DPA.

B. Move List Size

CBRS standards refer to the set of authorized transmissions that a SAS must suspend in order to protect a DPA from harmful interference on a given channel as the "move list" (i.e., transmissions that ultimately must be "moved" to a different channel). Because the move list depends on the aggregate interference of all transmissions in the affected area, and because multiple SASs may be managing these transmissions, the standards specify the algorithm all SASs must use to determine which transmissions must be placed on the move list.

We used the Wireless Innovation Forum's reference implementation of the standardized move list algorithm [6] to calculate the move list of each DPA; but in order to assess the tradeoff of limiting the CBSD antenna height, we replaced the ITM propagation model in the reference implementation with the hybrid propagation model for federal incumbent protection described in Section II-B.

An input to the move list algorithm are the sample points in the DPA at which the algorithm will ensure the aggregate interference meets the protection criteria. In this analysis, we used the default setting of the reference implementation, which is approximately 35 points along the edge of the DPA and approximately 15 points interior to the DPA.

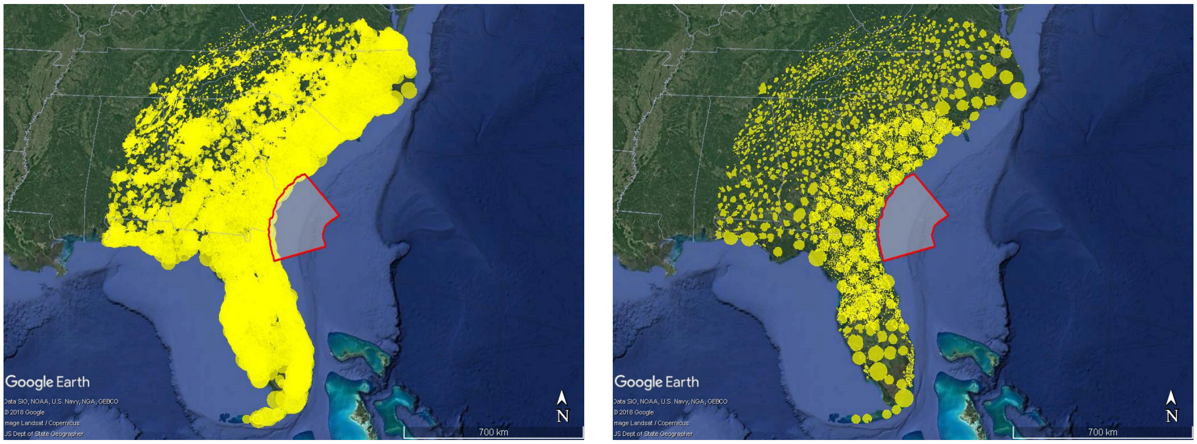
Fig. 4 illustrates the number of transmissions on the resulting move list for each DPA with and without the height restriction on a logarithmic scale. The dashed lines above each bar indicate the total number of CBSDs in the deployment. We observe different behavior depending on the DPA. For both West 14 and Pensacola, the move list increases modestly (about $5\,\%$) under the height restriction (despite the fact that the total number of CBSDs increases $16\,\%$ in the Pensacola deployment). However, though the East 2 deployment increases $8\,\%$ with the height restriction, the move list *decreases* by $25\,\%$. This result suggests that, in this case, the negative effect of the height restriction in terms of an increased number of required CBSDs is more than offset by its positive effect in the lower number of transmissions that are suspended in the event of incumbent activity.

C. DPA Neighborhood

The DPA neighborhood is the geographic area a SAS must consider when computing move lists. Any CBSD in the DPA neighborhood must be included in the computation and could potentially end up on the move list. The neighborhood is typically expressed in terms of a distance from the DPA boundary (or point) and is an input to the move list algorithm. The distance is such that transmissions beyond the neighborhood contribute negligible interference to the aggregate.

An estimate of the required neighborhood distance can be found by first computing the move list with artificially large neighborhood distances (in our case, 250 km for Category A CBSDs and 600 km for Category B CBSDs). The move list is composed of the union of component move lists, one for each protection point provided to the algorithm. For each transmission on the move list, we identify the component move

[1] We note that the sum of the areas of the contours in the original deployment is $6\,201\,917$ km^2, an order of magnitude larger than the area of their union, meaning that the original deployment has significant overlap of the -96 dBm/10 MHz contours.


Google Earth
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2018 Google
Image Landsat / Copernicus
US Dept of State Geographer
700 km
N
Google Earth
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
© 2018 Google
Image Landsat / Copernicus
US Dept of State Geographer
700 km
N

(a) Original Deployment (b) Non-Overlapping Subset

Fig. 2. −96 dBm/10 MHz contours of CBSDs deployed near the East 2 DPA.

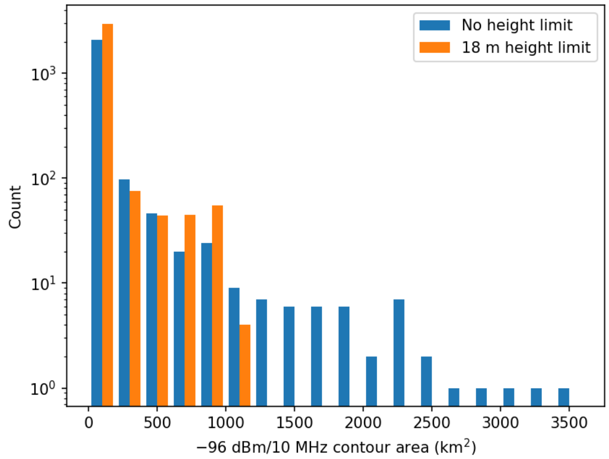


Fig. 3. Histogram of East 2 category B CBSD coverage by antenna height.

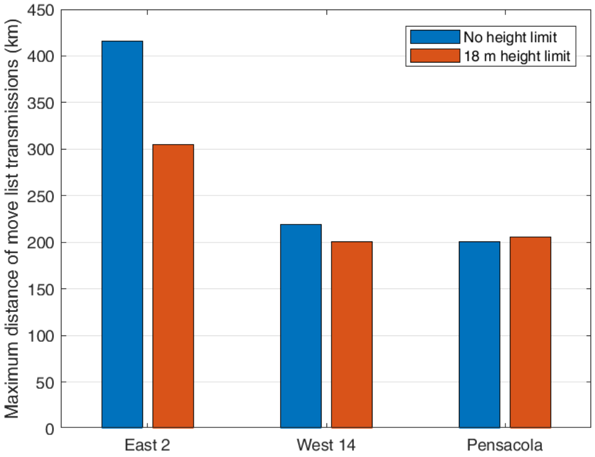


Fig. 5. Estimated Category B neighborhood distance.

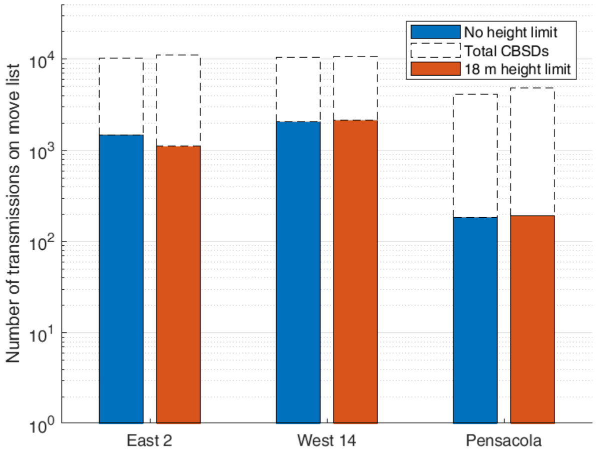


Fig. 4. Number of transmissions on the move list.

lists it belongs to and their associated protection points. We then find the distance from that transmission to the nearest of these protection points. Finally, from these distances, one per move-list transmission, we estimate the required neighborhood distance to be the largest of these distances.

In NTIA's initial release of DPA protection criteria [10], the East 2 DPA had the largest Category B neighborhood distance of all the DPAs, in excess of 400 km. The results of this analysis are consistent with that, as shown in Fig. 5 which plots the estimated Category B neighborhood distance of each DPA. However, our analysis shows that East 2 also stands to benefit the greatest from a height-restricted deployment when the hybrid propagation model is applied, lowering the neighborhood distance by more than 100 km. The change for the other two DPAs considered here, however, is small.

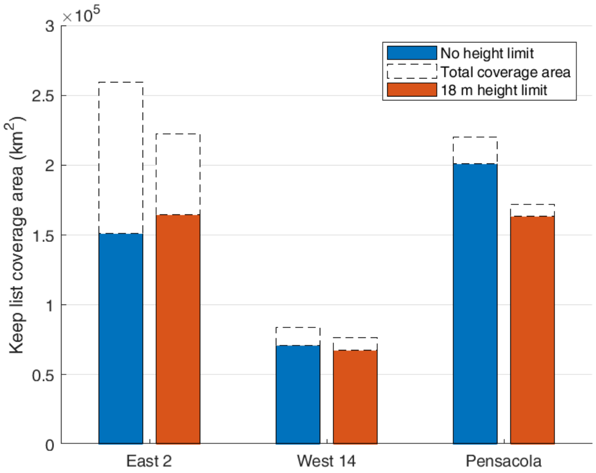
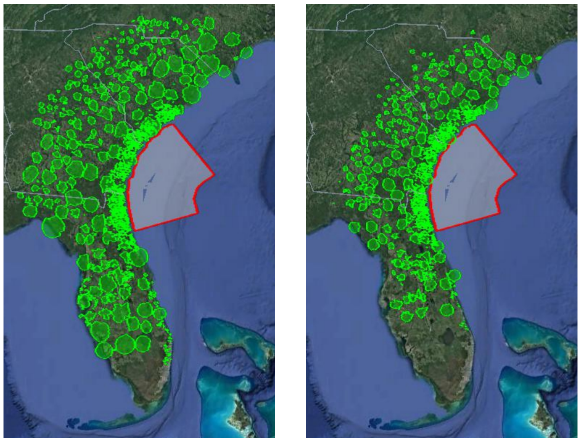


Fig. 6. Coverage area of the keep list.



(a) Original Antenna Heights (b) Restricted Antenna Heights

Fig. 7. −96 dBm/10 MHz contours of the East 2 DPA move list (map data: SIO, NOAA, U.S. Navy, NGA, GEBCO; map image: Landsat/Copernicus; copyright 2018 Google, INEGI).

D. Keep-List Coverage

The third and final metric used to evaluate the impact of incumbent activity on commercial usage is coverage area. In particular, we consider the coverage area of commercial service before and after a DPA becomes active.

The complement of the move list is referred to as the “keep list,” the list of transmissions that are not suspended by a SAS when a DPA is active. Fig. 6 shows the coverage area of all transmissions on the keep list for each DPA, with and without the height restriction. The dashed lines above each bar indicate the total coverage area when the DPA is inactive, that is, when no transmissions are suspended.

It is notable that, for East 2, though the total coverage area decreases with the height restriction, the keep-list coverage area actually *increases*, again pointing to the offsetting benefit of using lower CBSD antennas. For the other two DPAs, the keep-list coverage decreases with the height restriction, but not as much as the total area, as a percentage.

Fig. 7 shows the coverage contours of the transmissions on the East 2 move lists. The maps illustrate the smaller area affected by the move list for CBSDs with the height restriction relative to that without the height restriction.

V. Conclusion

The analysis quantified the effects of federal incumbent activity on commercial service in the 3.5 GHz CBRS band in terms of both the number of commercial transmissions that would be suspended as well as the affected service area. It also examined the tradeoff between commercial coverage and immunity to incumbent activity when the antenna heights of commercial base stations are capped.

Based on an analysis of a sample of federal incumbent protection areas, there are situations in which commercial users can benefit from limiting base station antenna heights. To realize this benefit, though, the propagation model used by the SAS for incumbent protection would need to be augmented with height-dependent clutter loss.

Future work can utilize the methodology presented here to examine the tradeoffs when changing other deployment variables, such as base station transmit power.

Disclaimer

Any mention of commercial products within this paper is for information only; it does not imply recommendation or endorsement by the authors.

References

[1] “Citizens broadband radio service,” 47 C.F.R. § 96, 2016.
[2] M. R. Souryal, T. T. Nguyen, and N. J. LaSorte, “3.5 GHz federal incumbent protection algorithms,” in *Proc. IEEE Dynamic Spectrum Access Networks (DySPAN)*, Oct. 2018.
[3] “Requirements for commercial operation in the U.S. 3550–3700 MHz citizens broadband radio service band,” Wireless Innovation Forum Document WINNF-TS-0112, Version V1.8.0, Jun. 2019. [Online]. Available: https://cbrs.wirelessinnovation.org/release-1-standards-specifications
[4] Irregular Terrain Model. [Online]. Available: https://www.its.bldrdoc.gov/resources/radio-propagation-software/itm/itm.aspx
[5] E. Drocella, J. Richards, R. Sole, F. Najmy, A. Lundy, and P. McKenna, “3.5 GHz exclusion zone analyses and methodology,” National Telecommunications and Information Administration, Technical Report TR 15-517, Mar. 2016. [Online]. Available: http://www.its.bldrdoc.gov/publications/2805.aspx
[6] SAS Testing and Interoperability Repository. [Online]. Available: https://github.com/Wireless-Innovation-Forum/Spectrum-Access-System
[7] SAS Data Repository. [Online]. Available: https://github.com/Wireless-Innovation-Forum/SAS-Data
[8] Letter from Paige R. Atkins, Assoc. Admin., Office of Spectrum Mgt., NTIA, to Julius P. Knapp, Chief, Office of Eng. and Tech., FCC, May 2018. [Online]. Available: https://www.fcc.gov/ecfs/filing/1051705880764
[9] “Operations for citizens broadband radio service; GAA spectrum coordination technical report–approach 3,” Wireless Innovation Forum Document WINNF-TR-2005, Version V1.0.0, May 2019. [Online]. Available: https://cbrs.wirelessinnovation.org/cbrs-reports-and-recommendations
[10] Dynamic Protection Area kml files, Oct. 2018. [Online]. Available: https://github.com/Wireless-Innovation-Forum/Spectrum-Access-System/tree/master/data/ntia

BowTie - a deep learning feedforward neural network for sentiment analysis[⋆]

Apostol Vassilev

National Institute of Standards and Technology,
100 Bureau Dr., Gaithersburg, MD 20899, USA
apostol.vassilev@nist.gov

Abstract. How to model and encode the semantics of human-written text and select the type of neural network to process it are not settled issues in sentiment analysis. Accuracy and transferability are critical issues in machine learning in general. These properties are closely related to the loss estimates for the trained model. I present a computationally-efficient and accurate feedforward neural network for sentiment prediction capable of maintaining low losses. When coupled with an effective semantics model of the text, it provides highly accurate models with low losses. Experimental results on representative benchmark datasets and comparisons to other methods[1]show the advantages of the new approach.

Keywords: Deep learning · Sentiment analysis · Natural Language Processing.

1 Introduction

When approaching the problem of applying deep learning to sentiment analysis one faces at least five classes of issues to resolve. First, what is the best way to encode the semantics in natural language text so that the resulting digital representation captures well the semantics in their entirety and in a way that can be processed reliably and efficiently by a neural network and result in a highly accurate model? This is a critically important question in machine learning because it directly impacts the viability of the chosen approach. There are multiple ways to encode sentences or text using neural networks, ranging from a simple encoding based on treating words as atomic units represented by their rank in a vocabulary [1], to using word embeddings or distributed representation of words [2], to using sentence embeddings. Each of these encoding types has different complexity and rate of success when applied to a variety of tasks.

[⋆] Supported by NIST Information Technology Laboratory Grant #7735282-000.

[1] DISCLAIMER: This paper is not subject to copyright in the United States. Commercial products are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

Vassilev, Apostol. "BowTie - a deep learning feedforward neural network for sentiment analysis." Paper presented at LOD 2019, Siena, IT. September 10, 2019 - September 13, 2019.

The simple encoding offers simplicity and robustness. The usefulness of word embeddings has been established in several application domains, but it is still an open question how much better it is than the simple encoding in capturing the entire semantics of the text in natural language processing (NLP) to provide higher prediction accuracy in sentiment analysis. Although intuitively one may think that because word embeddings do capture some of the semantics contained in the text this should help, the available empirical test evidence is inconclusive. Attempts to utilize sentence embeddings have been even less successful [3].

Second, given an encoding, what kind of neural network should be used? Some specific areas of applications of machine learning have an established leading network type. For example, convolutional neural networks are preferred in computer vision. However, because of the several different types of word and sentence encoding in natural language processing (NLP), there are multiple choices for neural network architectures, ranging from feedforward to convolutional to recurrent neural networks.

Third, what dataset should be used for training? In all cases the size of the training dataset is very important for the quality of training but the way the dataset is constructed and the amount of meta-data it includes also play a role. The Keras IMDB Movie reviews Dataset [4] (KID) for sentiment classification contains human-written movie reviews. A larger dataset of similar type is the Stanford Large Movie Review Dataset (SLMRD) [5]. I consider KID and SLMRD in Sections 2.3 and 2.4 respectively. Generally, simpler encodings and models trained on large amounts of data tend to outperform complex systems trained on smaller datasets [2].

Fourth, what kind of training procedure should be employed - supervised or unsupervised? Traditionally, NLP systems are trained on large unsupervised corpora and then applied on new data. However, researchers have been able to leverage the advantages of supervised learning and transfer trained models to new data by retaining the transfer accuracy [3].

Fifth, when training a model for transfer to other datasets, what are the model characterizing features that guarantee maintaining high/comparable transfer accuracy on the new dataset? Certainly, training and validation accuracy are important but so are the training and validation losses. Some researchers argue that the gradient descent method has an implicit bias that is not yet fully understood, especially in cases where there are multiple solutions that properly classify a given dataset [6]. Thus, it is important to have a neural network with low loss estimates for a trained model to hope for a good and reliable transfer accuracy.

The primary goal of this paper is to shed light on how to address these issues in practice. To do this, I introduce a new feedforward neural network for sentiment analysis and draw on the experiences from using it with two different types of word encoding: a simple one based on the word ranking in the dataset vocabulary; the other judiciously enhanced with meta-data related to word polarity. The main contribution of this paper is the design of the BowTie neural network in Section 3.

2 Data encoding and datasets

As discussed above, there are many different types of encodings of text with different complexity and degree of effectiveness. Since there is no convincing positive correlation established in the literature between complexity of the encoding and higher prediction accuracy, it is important to investigate the extent to which simple data encodings can be used for sentiment analysis. Simpler encodings have been shown to be robust and efficient [1]. But can they provide high prediction accuracy in sentiment analysis?

I investigate this open question by evaluating the accuracy one may attain using two types of text encoding on representative benchmark datasets.

2.1 Multi-hot encoding

The first encoding is the well-known *multi-hot* encoding of text [7]. The multi-hot encoding represents a very simple model of the semantics in text.

2.2 Polarity-weighted multi-hot encoding

The second encoding I consider is the polarity-weighted multi-hot encoding of text.

Let π be a linguistic type (e.g. morpheme, word) and let Π_D be the set of all such linguistic types in a dataset D. Let $M = |\Pi_D|$ be the cardinality of Π_D. Let ψ be a linguistic text type (e.g., a movie review) and let Ψ_D be the set of all texts in D. Let $N = |\Psi_D|$ be the cardinality of Ψ_D. Let Π_D and Ψ_D be finite sets such that the elements in each set are enumerated by $\{0, ..., M\}$ and $\{0, ..., N\}$ respectively. Let T^{NxM} be a tensor of real numbers of dimensions N by M, whose elements are weighted by the cumulative effect of the polarity of each word present in a given text π, as computed by [8]. Let $c_{\pi,\psi}$ be the number of tokens of the linguistic type π in a text ψ. Let ξ_π be the polarity rating of the token $\pi \in \Pi_D$. Naturally, I assume that if Ξ_D is the set of all polarity ratings for tokens $\pi \in \Pi_D$, then $|\Xi_D| = |\Pi_D|$. Let $\omega_{\xi\pi\psi} = \xi_\pi * c_{\pi,\psi}$ be the cumulative polarity of π in the text ψ. Let $\Omega_D = \{\omega_i\}_{i=0}^{M}$ and $C_D = \{c_i\}_{i=0}^{M}$. Let Θ^{NxM} be a tensor of real numbers of dimensions N by M, whose elements are set as follows:

$$\{\theta_{jk}\} = \begin{cases} \omega_{\xi_k \pi_k \psi_j}, & \text{if } \pi_k \in \psi_j; \\ \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The polarity-weighted multi-hot encoding (1) represents a more comprehensive model of the semantics in ψ, $\forall \psi \in \Psi_D$, that captures more information about ψ. I will attempt to investigate if and how much this additional information helps to improve the sentiment predictions in Section 4. An example of a polarity-weighted multi-hot encoded text is shown in Figure 1.

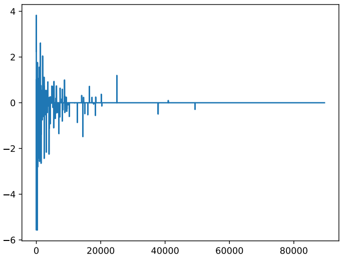


Fig. 1. A polarity-weighted multi-hot encoded text in D, Π_D with $M = 89527$.

2.3 The Keras IMDB Dataset (KID)

The KID [4] contains 50 000 human-written movie reviews that are split in equal subsets of 25 000 for training and testing and further into equal categories labeled as positive or negative. For convenience, the reviews have been pre-processed and each review is encoded as a sequence of integers, representing the ranking of the corresponding word in Π_D with $|\Pi_D| = 88\,587$. Hence, it can be easily encoded by the multi-hot encoding [7].

2.4 The Stanford Large Movie Review Dataset (SLMRD)

SLMRD contains 50 000 movie reviews, 25 000 of them for training and the rest for testing. The dataset comes also with a processed bag of words and a word polarity index [5, 10]. SLMRD contains also 50 000 unlabeled reviews intended for unsupervised learning. It comes with a Π_D, polarity ratings Ω_D, and word counts C_D with $|\Omega_D| = |C_D| = |\Pi_D| = 89\,527$.

3 The BowTie feedforward neural network

The ability of a neural network to provide accurate predictions and maintain low losses is very important for transferability to other datasets with the same or higher prediction accuracy as on the training dataset.

I now introduce a feedforward neural network with that criteria in mind. By way of background [11], logistic regression computes the probability of a binary output $\hat{y}_i$ given an input x_i as follows:

$$P(\hat{\mathbf{y}}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \mathbf{Ber}[\hat{y}_i|\mathbf{sigm}(x_i\mathbf{w})], \qquad (2)$$

where $\mathbf{Ber}[\ ]$ is the Bernouli distribution, $\mathbf{sigm}()$ is the sigmoid function, $\mathbf{w}$ is a vector of weights. The cost function to minimize is $\mathbf{C}(\mathbf{w}) = -\log P(\hat{\mathbf{y}}|\mathbf{X}, \mathbf{w})$.

This method is particularly suitable for sentiment prediction. One critical observation is that logistic regression can be seen as a special case of the generalized linear model. Hence, it is analogous to linear regression. In matrix form, linear regression can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}, \tag{3}$$

where $\hat{\mathbf{y}}$ is a vector of predicted values $\hat{y}_i$ that the model predicts for $\mathbf{y}$, $\mathbf{X}$ is a matrix of row vectors x_i called regressors, $\mathbf{w}$ are the regression weights, and $\boldsymbol{\epsilon}$ is an error that captures all factors that may influence $\hat{\mathbf{y}}$ other than the regressors $\mathbf{X}$.

The gradient descent algorithm used for solving such problems [11] may be written as

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \rho^{(k)}\mathbf{g}^{(k)} + \boldsymbol{\epsilon}^{(k)}, \tag{4}$$

where $\mathbf{g}^{(k)}$ is the gradient of the cost function $\mathbf{C}(\mathbf{w})$, $\rho^{(k)}$ is the learning rate or step size, and $\boldsymbol{\epsilon}^{(k)}$ is the error at step k of the iterative process. One error-introducing factor in particular is the numerical model itself and the errors generated and propagated by the gradient descent iterations with poorly conditioned matrices running on a computer with limited-precision floating-point numbers. Even if regularization is used, the specific parameters used to weigh them in the equation, e.g. the L_2-term weight or the dropout rate, may not be optimal in practice thus leading to potentially higher numerical error. This is why it is important to look for numerical techniques that can reduce the numerical error effectively. This observation inspires searching for techniques similar to multigrid from numerical analysis that are very effective at reducing the numerical error [12].

The neural network design

The feedforward neural network at the bottom of Figure 2 consists of one encoding layer, a cascade of dense linear layers with L_2-regularizers and of appropriate output size ($Z_i@K$, where K is the layer output size) followed by a dropout regularizer and a sigmoid. The encoder takes care of encoding the input data for processing by the neural network. In this paper I experiment with the two encodings defined in Section 2: the simple multi-hot encoding and the polarity-weighted multi-hot encoding.

The sigmoid produces the estimated output probability $P(\hat{\mathbf{y}}^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}^{(i)})$, which may be used to compute the negative log-loss or binary cross-entropy as

$$-\left[\mathbf{y}\log(P(\hat{\mathbf{y}}^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}^{(i)})) + (1-\mathbf{y})\log(1 - P(\hat{\mathbf{y}}^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}^{(i)}))\right]. \tag{5}$$

The binary cross-entropy provides a measure for quality and robustness of the computed model. If the model predicts correct results with higher probability then the binary cross-entropy tends to be lower. If, however, the model predicts correct results with probability close to the discriminator value or predicts an incorrect category, the binary cross-entropy tends to be high. Naturally, it is

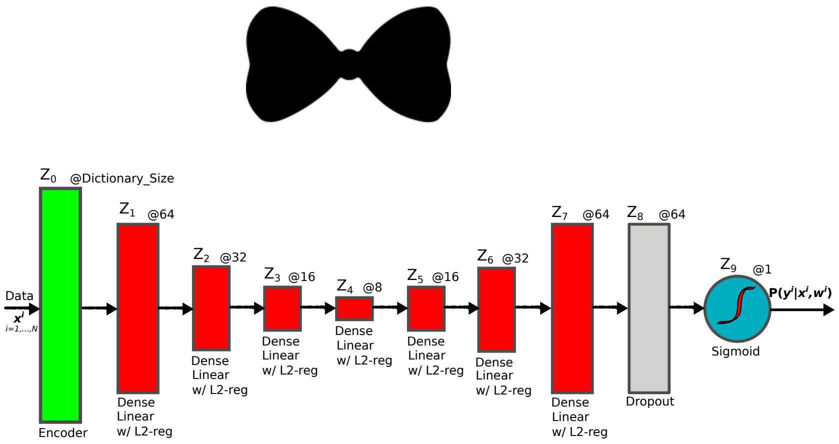


Fig. 2. The BowTie neural network. The classic bow tie (shown on top) originated among Croatian mercenaries during the Thirty Years' War of the 17th century. It was soon adopted by the upper classes in France, then a leader in fashion, and flourished in the 18th and 19th centuries. (Wikipedia). The estimated probability $P(\hat{\mathbf{y}}^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}^{(i)})$ may be fed into a post-processing discriminator component to assign a category (pos/neg) for the input $\mathbf{x}^{(i)}$ with respect to a discriminator value $\delta \in [0, 1]$. All experiments presented in this paper use $\delta = 0.5$.

Table 1. BowTie hyperparameters.

Hyperparameters	
name	**values/range**
L_2-regularization weight	0.01 - 0.02
Dropout rate	0.2 - 0.5
Optimizer	NADAM or RMSProp
Dense Layer Activation	None (Linear network), RELU

desirable to have models that confidently predict correct results. It is also important to have models that maintain low binary cross-entropy for many training epochs because, depending on the training dataset, the iterative process (4) may need several steps to reach a desired validation accuracy. Models that quickly accumulate high cross-entropy estimates tend to overfit the training data and do poorly on the validation data and on new datasets.

Hyperparameters. There are several hyperparameters that influence the behavior of BowTie, see Table 1. For optimal performance the choice of the dense layer activation should be coordinated with the choice for the L_2-regularization weight. This recommendation is based on the computational experience with

BowTie and is in line with the findings in [14] about the impact of the activation layer on the training of neural networks in general. The optimal L_2-regularization weight for the linear network (no dense layer activation) is 0.019 whereas rectified linear unit (RELU) activation favors L_2-regularization weights close to 0.01. The network can tolerate a range of dropout rates but good stability and accuracy is attained with a dropout rate of 0.2. It is interesting to note thatRoot Mean Square Propagation (RMSProp) tends to converge faster to a solution and sometimes with a higher validation accuracy than Nesterov adaptive momentum (NADAM) but the transfer accuracy of the models computed with NADAM tends to be higher than for models computed with RMSPRop. For example, a model trained on SLMRD with validation accuracy of 89.24 %, higher than any of the data in Table 4 below, yielded 91.04 % transfer accuracy over KID, which is lower than the results in Table 4. This experimental finding is consistent over many tests with the two optimizers and needs further investigation in future research to explore the theoretical basis for it.

It may be possible to run BowTie with other optimizers or activation layer choices than those shown in Table 1 but I have not tested their effectiveness.

4 Training and transfer scenarios

This section defines the objectives for the testing of the BowTie neural network shown in Figure 2 in terms of four training and transfer scenarios.

But first it is important to decide on the type of training to employ - supervised or unsupervised. I embark on supervised training based on the findings in [3] about the advantages of supervised training and the availability of corpora of large labeled benchmark datasets [4] and [5]. The scenarios to explore are:

– **Scenario 1 (Train and validate):** Explore the accuracy and robustness of the BowTie neural network with the simple multi-hot encoding by training and validating on KID.
– **Scenario 2 (Train and validate):** Explore the accuracy and robustness of the BowTie neural network with the simple multi-hot encoding by training and validating on SLMRD.
– **Scenario 3 (Train and validate):** Explore the accuracy and robustness of the BowTie neural network with the polarity-weighted multi-hot encoding by training and validating on SLMRD.
– **Scenario 4 (Train, validate, and transfer):** Explore the transfer accuracy of the BowTie neural network with polarity-weighted multi-hot encoding by training on SLMRD and predicting on KID.

The primary goal of this exploration is to establish some baseline ratings of the properties of the BowTie neural network with the different encodings and compare against similar results for other neural networks with other types of encoding. This provides a quantitative criteria for comparative judging.

Results

In this section I report the results from executing Scenarios 1-4 from Section 4 using TensorFlow [7], version 1.12, on a 2017 MacBook Pro, *without* GPU acceleration. The modeling code is written in Python 3. The speed of computation improved on a platform *with* eight Tesla V100-SXM2 GPUs: speedup of nearly a factor of two but in training the acceleration plateaued if more than two GPUs were used.

Scenario 1. In this test, the BowTie neural network is tested with encoding [7]. The results in Table 2 show high accuracy and low binary cross-entropy estimates.

Table 2. Scenario 1 results. Data from experiments with training a model on KID until it attains some validation accuracy grater than 88 %. Note that each time the data is loaded for training, it is shuffled randomly, hence the small variation in computational results.

Training and validating on KID	
validation accuracy (%)	validation binary cross-entropy
88.08	0.2955
88.18	0.2887
88.21	0.2945

To assess the relative computational efficiency of the BowTie neural network, I compared it to the convolutional neural network in [13] with a 10000 word dictionary. The network reached accuracy of 88.92 % at Epoch 4 with binary cross-entropy of 0.2682. However, it took 91 seconds/Epoch, which matched the numbers reported by the authors for the CPU-only computational platform [13]. In addition, after Epoch 4, the binary cross-entropy started to increase steadily while the accuracy started to decline. For example, the binary cross-entropy reached 0.4325 at Epoch 10 with validation accuracy of 87.76 % and 0.5536 and 87.40 % correspondingly at Epoch 15. In comparison, BowTie takes only 3 seconds/Epoch for the same dictionary size and attains accuracy of 88.20 % with binary cross-entropy of 0.2898. The binary cross-entropy stays below 0.38 for a large number of Epochs.

Scenario 2. SLMRD is more challenging than KID for reasons that are visible in the test results for Scenarios 3 and 4, hence the slightly lower validation accuracy attained by BowTie using the simple multihot encoding [7] - it easily meets or exceeds the threshold accuracy of 87.95 % but could not surpass the 88 % level in several experiments.

Table 3. Scenario 2 results. Data from experiments with training a model on KID until it attains validation accuracy of at least 87.95 %. Note that each time the data is loaded for training, it is shuffled randomly, hence the small variation in computational results.

Training and validating on SLMRD	
validation accuracy (%)	validation binary cross-entropy
87.98	0.2959
87.95	0.2996
87.96	0.3001

Scenarios 3 and 4. I combine the reporting for Scenarios 3 and 4 because once the model is trained under Scenario 3 it is then transferred to compute predictions on KID. To perform the transfer testing on KID one needs to reconcile the difference in $|\mathit{\Pi}_{SLMRD}|$ and $|\mathit{\Pi}_{KID}|$. As I noted in Section 2, $|\mathit{\Pi}_{SLMRD}| = 89\,527$ and $|\mathit{\Pi}_{KID}| = 88\,587$. Moreover, $\mathit{\Pi}_{KID} \not\subset \mathit{\Pi}_{SLMRD}$. Let $\mathit{\Pi}_{\Delta} = \mathit{\Pi}_{KID} \setminus (\mathit{\Pi}_{KID} \cap \mathit{\Pi}_{SLMRD})$. It turns out that

$$\mathit{\Pi}_{\Delta} = \left\{ \begin{array}{l} 0s, 1990s, 5, 18th, 80s, 90s, 2006, 2008, \\ 85, 86, 0, 5, 10, tri, 25, 4, 40s, 70s, 1975, \\ 1981, 1984, 1995, 2007, dah, walmington, \\ 19, 40s, 12, 1938, 1998, 2, 1940's, 3, 000, 15, 50. \end{array} \right\}.$$

Clearly, $|\mathit{\Pi}_{\Delta}|$ is small and of the words in $\mathit{\Pi}_{\Delta}$, only 'walmington' looks like a plausible English word. This is the name of a fictional town in a British TV series from the 1970's. As such it has no computable polarity index and is semantically negligible. Based on this, I dropped all these words during the mapping of $\pi \in \mathit{\Pi}_{\Delta}$ into the corresponding $\pi\prime \in \mathit{\Pi}_{SLRMD}$. Note that the mapping $\pi \rightarrow \pi\prime$ enables the encoding of the semantics of the texts in KID according to (1).

Table 4. Scenarios 3 and 4 results. Data from experiments with training a model on SLMRD until it attains validation accuracy grater than 89 % and using that model to predict the category for each labeled review in KID, computing over the entire set of 50 000 reviews in KID.

Training on SLMRD		**Predicting on KID**
validation accuracy (%)	val. binary cross-entropy	Tansfer accuracy (%)
89.02	0.2791	91.56
89.12	0.2815	91.63
89.17	0.2772	91.76

Some simple but revealing statistics about the data. With encoding (1), the elements of the matrix T_{SLMRD} for the training set in SLMRD show cumulative polarity values in the range [-50.072 837, 58.753 546] and the cumulative polarity of the elements of the matrix T_{SLMRD} for the test set is in the range [-48.960 107, 63.346 164]. This suggests that SLMRD is pretty neutral and an excellent dataset for training models. In contrast, the elements of the matrix T_{KID} are in the range [-49.500 000, 197.842 862]. Table 4 contains the results from executing Scenarios 3 and 4. The validation and transfer accuracy results in Table 4 are better than those shown in Tables 2 and 3. This demonstrates the value word polarity brings in capturing the semantics of the text. The transfer accuracy over KID is also higher than the results obtained by using the convolutional neural network [13] on KID. The results in Table 4 are higher than the results reported for sentiment prediction of movie reviews in [3] but in agreement with the reported experience by these authors about consistently improved accuracy from supervised learning on a large representative corpus before transferring the trained model to a corpus of interest.

Note also that the validation accuracy of BowTie with the encoding (1) on SLMRD is higher than the results in [10] for the same dataset. The observation in [10] that even small percentage improvements on a large corpus result in a significant number of correctly classified reviews applies to the data in Table 4: that is, there are between 172 and 210 more correctly classified reviews by BowTie.

5 Discussion and next steps

The experimental results from sentiment prediction presented above show the great potential deep learning has to enable automation in areas previously considered impossible. At the same time, we are witnessing troubling trends of deterioration in cybersecurity that have permeated the business and home environments: people often cannot access the tools they need to work or lose the data that is important to them [15]. Traditionally, governments and industry groups have approached this problem by establishing security testing and validation programs whose purpose is to identify and eliminate security flaws in IT products before adopting them for use. One area of specific concern in cybersecurity is cryptography. Society recognizes cryptography's fundamental role in protecting sensitive information from unauthorized disclosure or modification. The cybersecurity recommendations in [15] list relying on cryptography as a means of data protection as one of the top recommendations to the business community for the past several years. However, the validation programs to this day remain heavily based on human activities involving reading and assessing human-written documents in the form of technical essays. This validation model worked well for the level of the technology available at the time when the programs were created more than two decades ago. As technology has advanced, however, this model no longer satisfies current day industry and government operational needs in the context of increased number and intensity of cybersecurity breaches [15].

There are several factors for this. First, current cybersecurity recommendations [15] require patching promptly, including applying patches to cryptographic modules. Technology products are very complex and the cost of testing them fully to guarantee trouble-free use is prohibitively high. As a result, products contain vulnerabilities that hackers and technology producers are competing to discover first: the companies to fix, the hackers to exploit. Patching products changes the game for hackers and slows down their progress. However, patching changes also the environment in which a cryptographic module runs and may also change the module itself, thus invalidating the previously validated configuration. Users who depend on validated cryptography face a dilemma when frequent updates and patches are important for staying ahead of the attackers, but the existing validation process does not permit rapid implementation of these updates while maintaining a validated status because of the slow human-based validation activities.

The second factor hindering the effectiveness of the traditional validation model is the demand for speed in the context of the cognitive abilities of the human brain. Recent scientific research points out that humans are limited in their ability to process quickly and objectively large amounts of complex data [16].

Going back to the results on sentiment analysis with deep learning from above and in spite of that success, people may always question the ability of machines to replace humans in solving such cognitive and analytical tasks. They will always ask why the accuracy is not one hundred percent? Or, notwithstanding the available scientific evidence [16], say that if a human was reviewing the text she would have never made a mistake.

Changing public opinion may be a slow process. Besides, it seems that cybersecurity and machine learning/artificial intelligence (AI) will always be joined at the hip because the more we rely on machines to solve ever more complex tasks, the higher the risk those machines may be attacked to subvert their operation. Can AI fight back though? The results presented in this paper suggest that computer-based validation of cryptographic test evidence may be the only viable alternative that would allow objective and accurate assessment of large volumes of data at the speed needed to respond to the present day cybersecurity challenge [15]. This paper demonstrates that deep learning neural networks are capable of tackling the core tasks in security validation and thereby automating existing programs [9]. If this effort is indeed successful one may reason that by helping to improve the process of validation and thereby increase cybersecurity, albeit indirectly, AI will in fact be defending itself from cybersecurity threats. Over time, this may lead to societal acceptance of AI into sensitive domains such as the validation of critical components for the IT infrastructure.

Next steps. The importance of incorporating word polarity into the model is illustrated clearly by the results presented in this paper. However, the type of language used in the validation test reports tends to be different than the colloquial English used in movie reviews. The technical jargon in test reports uses many common words whose meaning changes in this context. Moreover, the assessments for the test requirements are written in a way different from movie reviews. Here the author provides arguments that justify her conclusion about compliance. The challenge is to distinguish weak/faulty arguments from solid ones and develop an appropriate polarity model, along with assembling a representative corpus of labeled validation test report data for training and validation.

References

1. Brants, T., Popat, A., Xu, P., Och, F., Dean, J.: Large Language Models in Machine Translation. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (Prague), pp. 858-867.
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality, pp. 3111-3119. Advances in neural information processing systems (2013)
3. Conneau, A., Kiela, D., Schwenk, H., Barraul, L., Bordes, A.: Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (Copenhagen), pp. 670–680, Association of Computational Linguistics (2017)
4. IMDB Movie reviews sentiment classification, Keras Documentation, https://keras.io/datasets/. Last accessed 1 March 2019
5. Large Movie Review Dataset, Stanford University, http://ai.stanford.edu/ãmaas/data/sentiment/. Last accessed 1 March 2019
6. Nacson, M., Lee. J., Gunasekar, S., Savarese, P., Srebro, N., Soudry, D.: Convergence of Gradient Descent on Separable Data, https://arxiv.org/abs/1803.01905v2. Last accessed 15 March 2019
7. TensorFlow: An open source machine learning framework for everyone, Google LLC, https://www.tensorflow.org/. Last accessed 1 March 2019
8. Potts, C.: On the negativity of negation. In: Proceedings of Semantics and Linguistic Theory. vol. 20, pp. 636-659. (201‘)
9. Automated Cryptographic Validation Testing, NIST, https://csrc.nist.gov/projects/acvt/. Last accessed 1 March 2019
10. Maas, Andrew L., Daly, Raymond E., Pham, Peter T., Huang, Dan, Ng, Andrew Y., Potts, C.: Learning Word Vectors for Sentiment Analysis. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, Portland, Oregon, USA, pp. 142–150. Association for Computational Linguistics.
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016).
12. Bramble, J.: Multigrid Methods. Pitman Research Notes in Mathematics, Vol. 294, Wiley (1993), ISBN 0-470233524.
13. Demonstration of the use of Convolution1D for text classification, Keras, https://github.com/keras-team/keras/blob/master/examples/imdb_cnn.py. Last accessed 1 February 2019
14. Hayou, S., Doucet, A., Rousseau, J.: On the Impact of the Activation Function on Deep Neural Networks Training, https://arxiv.org/abs/1902.06853. Last accessed 1 March 2019
15. 2018 Data Breach Investigations Report, Verizon, https://enterprise.verizon.com/resources/reports/dbir/. Last accessed 1 February 2019
16. Kahneman, D.: Thinking, fast and slow. Farrar, Straus and Giroux, New York (2011), ISBN 9780374275631 0374275637.

A Science Gateway for Atomic and Molecular Physics

Barry I. Schneider
barry.schneider@nist.gov
National Institute of Standards and Technology, Gaithersburg, Maryland 20899, USA

Klaus Bartschat
Oleg Zatsarinny
klaus.bartschat@drake.edu
oleg_zoi@yahoo.com
Drake University, Des Moines, IA 50311, USA

Igor Bray
igor.bray@curtin.edu.au
Curtin University, Perth, GPO Box U1987, Western Australia

Armin Scrinzi
armin.scrinzi@lmu.de
Ludwig-Maximilians Universität, München, Germany

Fernando Martín
Markus Klinker
fernando.martin@uam.es
markusklinker@gmail.com
Universidad Autónoma de Madrid, Catoblanco, Madrid 28049 Spain

Jonathan Tennyson
j.tennyson@ucl.ac.uk
Department of Physics and Astronomy, University College London, Gower Street, London, WC1E 6BT United Kingdom

Jimena D. Gorfinkiel
jimena.gorfinkiel@open.ac.uk
School of Physical Sciences, The Open University, Milton Keynes MK7 6AA, United Kingdom

Sudhakar Pamidighantam
pamidigs@iu.edu
Indiana University, CIB 2709 E 10th Street, Bloomington, In 47408 and the eXtreme Science and Engineering Discovery Environment(XSEDE)

ABSTRACT

We describe the creation of a new Atomic and Molecular Physics science gateway (AMPGateway). The gateway is designed to bring together a subset of the AMP community to work collectively to make their codes available and easier to use by the partners as well as others. By necessity, a project such as this requires the developers to work on issues of portability, documentation, ease of input, as well as making sure the codes can run on a variety of architectures. Here we outline our efforts to build this AMP gateway and future directions.

KEYWORDS

Atomic and Molecular physics, Science Gateway, Light matter interaction, Ab initio Quantum Physics

ACM Reference format:
Barry I. Schneider, Klaus Bartschat, Oleg Zatsarinny, Igor Bray, Armin Scrinzi, Fernando Martín, Markus Klinker, Jonathan Tennyson, Jimena D. Gorfinkiel, and Sudhakar Pamidighantam. 2019. A Science Gateway for Atomic and Molecular Physics. In *Proceedings of PEARC '19: Practice and Experience in Advanced Research Computing: Rise of the Machines, Chicago, IL, July 28 - August 1, 2019 (PEARC '19),* 8 pages.
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

On May 14-16, 2018, an NSF supported workshop entitled, “Developing Flexible and Robust Software in Computational Atomic and Molecular (A&M) Physics” was organized by Barry Schneider (chair), Robert Forrey (Penn State) and Naduvalath Balakrishnan (UNLV) at the Institute for Theoretical Atomic and Molecular Physics, Harvard-Smithsonian ITAMP [5]. The purpose of the workshop was to bring together a group of internationally known researchers in computational atomic and molecular physics to:

- Identify and prioritize outstanding problems in A&M science, which would benefit from a concerted community effort in developing new software tools and algorithms that would lead to more rapid and productive scientific progress for the entire community.
- Discuss approaches to optimize achieving that goal.
- Produce and disseminate a report of the workshop to the community.

A concerted community effort is underway to develop and maintain these tools in order to ensure continued scientific progress. The group acknowledged that, in contrast to some other communities, A&M physics has lagged behind in developing community software that is robust and can be used, in a relatively straightforward way, by other than the group who developed that software. While there are exceptions, many software packages are poorly documented, poorly written, and only usable by a set of local “experts”. The tools themselves are capable of treating scientific and technologically interesting problems, but they are only accessible to a small group of people. The codes are not always maintained, and the lack of coordination among the developers has led to a lot of “reinventing the wheel”. The group felt strongly that the efforts being expended

Schneider, Barry I.; Bartschat, Klaus; Zatsarinny, Oleg; Bray, Igor; Martin, Fernando; Scrinzi, Armin; Pamidighantam, Sudhakar; Tennyson, Jonathan; Gorfinkiel, Jimena; Klinker, Markus. "A Science Gateway for Atomic and Molecular Physics." Paper presented at Practice and Experience in Advanced Research Computing (PEARC), Chicago, IL, US. July 28, 2019 - August 01, 2019.

in developing these computational tools should be available and usable by future generations of A&M scientists.

The success of the workshop led six of the groups to work together and develop an XSEDE proposal to build and maintain a Science Gateway devoted to the codes developed in these groups. That proposal was supported, and since May of 2018 there has been decent progress. A number of the codes are already ported and running on various XSEDE platforms. Some progress has been made in making them usable by others within the group but not yet the outside world. We are now taking steps to achieve this last goal.

The AMPGateway uses the multi-tenanted Apache Airavata middleware framework [2–4] served by the SciGaP hosting services for sustained operation. In the first stage of our efforts the software suites have been deployed as independent applications with specific input interfaces. Community building has already started and a few additional software suites have been identified for inclusion in phase two. The interoperability of the software suites is very important and will be addressed as a follow-on.

The present manuscript is divided into four major sections. In the Introduction we provide a history of how and why the project got started and our decision to go to XSEDE [1] for support for the gateway. In Section II we present some information on the AMP codes that are already available on the gateway. Section III is devoted to the details of the construction and deployment of the gateway. In Section IV we discuss issues of broadening usage of the gateway and questions of community building.

2 CURRENT CODE STATUS

At present, we have concentrated our major effort on five codes. A brief description of these packages is given below.

2.1 BSR

The *B*-spline *R*-matrix (BSR) method and the accompanying computer code [6] were developed by Oleg Zatsarinny in the group of Klaus Bartschat at Drake University. The program computes transition-matrix elements for electron collisions with atoms and ions as well as photoionization processes. From these, cross sections and other experimentally observable parameters can be obtained. The code can also be run in a mode that provides atomic structure information through energy levels and oscillator strengths.

The BSR approach is a particular variant of the *R*-matrix method to solve the close-coupling equations in coordinate space. In this respect, it is complementary to the convergent close-coupling (CCC) approach described below. BSR is an alternative formulation of the well-known *R*-matrix code developed in Belfast under the long-term leadership of Philip Burke. The Belfast code is somewhat singular in that it is readily available and used by a small group of users. While the last general write-up appeared in 1995 [7], updated versions are available [8]. A comprehensive introduction to *R*-matrix theory for atomic and molecular collisions processes, as well an overview of many applications, can be found in the book by Burke [9].

The published BSR code [6] is a serial version, which was written in the non-relativistic and semi-relativistic (Breit-Pauli) frameworks. Relativistic (DBSR) and MPI-parallelized versions, as well as extensions to treat ionization processes (similar to the CCC method described below) exist and are being used by the developer and a small group of collaborators. Executables of the parallelized codes (currently running on Stampede2) will be uploaded to the Gateway in the near future. The BSR and DBSR packages are a prime example where updated documentation and a wide distribution are urgently needed before critical expertise is lost. Fortunately, the urgency was recently recognized by the NSF and resulted in the funding of a three-year proposal to achieve exactly these goals. We expect the gateway described in the present paper to be one of the vehicles to ensure significant future progress.

A comprehensive overview of the BSR method and its applications at the time was published by Zatsarinny and Bartschat [10]. The most noteworthy features of the code are:

- Use a finite-element (*B*-spline) rather than a finite-difference approach in the calculation of the matrix elements needed to set up the hamiltonian in the inner region.
- Employ non-orthogonal sets of one-electron orbitals to account for the term-dependence of the valence orbitals, in particular for complex, open-shell targets, thereby providing an economical and accurate description of the target states and much flexibility in building the scattering wavefunction as well as pseudostates to further improve the target description and enable the treatment of electron-impact single-ionization as well as photon-driven double ionization processes.

The BSR code has the following major parts:

- Build the $N-$ and $(N+1)-$electron configurations.
- Generate all necessary one-electron and two-electron matrix elements to set up the target and scattering Hamiltonians in the internal region.
- Diagonalize these Hamiltonians.
- Propagate the wavefunction from the *R*-matrix boundary, $r = a$, to "asymptotia" (r_b), where it can be matched to known analytic forms. The propagation requires the solution of a set of coupled differential equations using known long-range potentials and needs to be repeated for each scattering energy. If angle-differential ionization processes with two free electrons in the final state are to be treated as well, the inner region may need to be increased beyond the original criterion.

Even though there is no general way to predict where most of the computational effort is needed, in most cases the generalized eigenvalue problem (diagonalization with *all* eigenvalues needed) of the $(N+1)-$electron Hamiltonian is a very time-consuming step. For complex targets, setting up this Hamiltonian can be expensive as well. For ionic targets, the wealth of resonances may require many thousands of collision energies to be treated, which can result in significant time going into the asymptotic region.

To summarize: The BSR method is closely related to both the Belfast *R*-matrix approach and the CCC method described below. The two *R*-matrix codes were designed to handle complex targets and many energies, while the CCC code can handle more processes but is essentially limited to quasi-one and quasi-two electron targets. Some benchmark comparisons for problems that all three methods should be able to handle have been performed. As expected, the results are numerically equivalent, but one or the other method may be more efficient. The details strongly depend on the complexity of the target and the energies for which results are required.

2.2 CCC

The original implementation of the Convergent Close-Coupling method was designed to produce accurate cross sections for scattering of light projectiles from quasi one- and two-electron targets [14]. It began with electron-hydrogen scattering [15], and was further extended to quasi-one electron targets well-modeled by one valence electron above a frozen Hartree-Fock core [16]. It was then extended to the helium target [17], and quasi two-electron targets such as Be [18]. The main features are:

- Expansion of the target state in a sufficiently complete $\mathscr{L}^2$ basis size N to treat cases where both excitation and ionization of the target are possible, with convergence tested by systematically increasing N.
- Expansion of the scattering wavefunction in the momentum based Lippmann-Schwinger equation.
- Introduction of numerical quadrature to reduce the problem to a very large set of linear algebraic equations.

CCC has been extended to positron scattering, where the positron introduces a second center capable of forming the electron positron bound state known as positronium (Ps) [19]. This is an example of a rearrangement collision and as such introduces even more complexity into the computational procedure. Such calculations can be "time-reversed" to be considered as Ps scattering on (anti)protons to form (anti)hydrogen [20]. A review of the CCC method for positron scattering has been given by Kadyrov and Bray [21].

On the computational side, the CCC codes have been parallelized to use OpenMP on a node and MPI between nodes and have been deployed successfully on Comet and Stampede2. A GPU implementation, currently underway, shows immense promise with up to two orders of magnitude speedup.

2.3 UKRMol(+)

The UK Molecular R-matrix codes were designed to treat low-energy elastic and inelastic electron-molecule collisions using the R-matrix method; they have evolved to also study photoionization and positron-molecule collisions as well as to produce the input required for time-dependent molecular R-matrix with time (RMT) calculations [22]. Similar to BSR they are based on the R-matrix method and the general theory can be found in the book by Burke [9].

The now frozen release version of the (mainly serial) UKRMol suite uses Gaussian Type Orbitals (GTOs) to represent both the target and continuum orbitals. A publication presenting this code by Carr *et al.* [23] followed a project which substantially updated (to Fortran95) and standardized the programming used, particularly in the inner region. A review article by Tennyson [25] from the same period gives a comprehensive overview of theory used and the functionality of the code.

The use of GTOs to represent the continuum leads to constraints on both the size of the inner region that can be used and the free electron energy range. Recently a new suite, known as UKRMol+, has been developed led by Zdeněk Mašín and Jimena Gorfinkiel [23]. The code uses the new GBTOlib integral library to determine all the one- and two-electron integrals needed in the mixed basis of GTOs and B-splines: it offers the choice of using GTOs, B-splines or hybrid GTOs – B-splines to represent the continuum; the bound orbitals are always described by GTOs. The library, that uses object oriented features from the Fortran2003 standard, involves distributed and shared-memory parallelization. UKRMol+ incorporates a number of algorithmic improvements, including a faster configuration state function (CSF) generation and parallization of the construction and diagonalization of the N and $(N+1)$ Hamiltonians [29]. Further parallelization to avoid I/O to disk during the evaulation of transition moments for photoionization and RMT input is currently being tested.

Both suites contain a rich array of outer region functionality including automated resonance detection and fitting, bound state detection, computation of multichannel quantum defects, rotational excitation and evaluation of photoionization cross sections. So far applications of the UKRMol+ suite are limited [26–28], but a full release and associated article will be available shortly [29].

The codes have been available as freeware for more than a decade and are widely used: the software can be downloaded as a tarball and installed (in the case of UKRmol+) using cmake, provided the necessary libraries are available in the system. Neither suite is straightforward to use without training; Quantemol-N [30] is a commercial front end which has led to a significant increase in the user base of the code. A set of perl scripts developed by Karel Houfek that simplify writing the input is also now available, both for electron scattering and photoionization calculations. Further details can be found on the website of the UK Atomic, Molecular and Optical physics R-matrix consortium (https://www.ukamor.com/).

2.4 tRecX

The tRecX code package [31, 32] is a general framework for solving initial value problems of the form

$$\frac{\partial}{\partial t}\Psi = \mathcal{D}[\Psi, t] + \Phi(t) \tag{1}$$

for an arbitrary number of spatial dimensions and a variety of coordinate systems. The main design is for linear $\mathcal{D}$, but non-linear operators can also be used.

2.4.1 Applications. The code has been primarily used for solving the time-dependent Schrödinger equation of atomic and molecular systems in ultra-short pulses and in strong near-IR fields. The most significant results are fully differential spectra for single- and double photo-emission from the He atom at near infrared wave-length [33], strong field ionization rates of noble gases [34] and differential spectra for small di- and tri-atomic molecules [35, 36], cf. Fig. 1.

2.4.2 Methods. The code uses three newly developed key techniques:

(1) irECS — *"Infinite range exterior complex scaling"* [37] as absorbing boundary conditions and for the computation of life-times. irECS is a variant of exterior complex scaling with an infinitely wide boundary for absorption at all wave lengths.

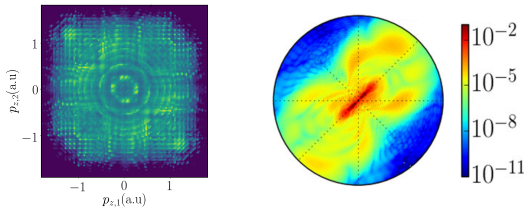


Figure 1: Left: Helium in full dimensions, double emission cross-section $\sigma(p_{z,1}, p_{z,2})$ for a 2-cycle pulse at wavelength 400 nm and $5{\times}10^{14} W/cm^2$ intensity. Anti-correlated emission is favored. Right: haCC calculation for CO_2, photoemission by an 800 nm laser pulse at intensity $10^{14} W/cm^2$ up to energies 2.5 au in the xz-plane at 45° alignment of the molecular axis to polarization direction (from [36].)

(2) tSurff — the *time-dependent surface flux* method [38] for photo-emission spectra. By tSurff, the actual numerical solution remains contained in a small reactive region of typically 20 to 100 atomic units.

(3) haCC — the *"hybrid anti-symmetrized Coupled Channels"* method [39] combines Gaussian-based neutral and ionic CI states with a numerical single-electron basis. The numerical basis extends over the whole system, thus ensuring the proper description of energetic electron-molecule collisions.

Two of these techniques are reflected in the code's name

tRecX = tSurff + irECS

2.4.3 Structure and inputs. An effort was made to keep the code flexible without sacrificing efficiency. Systems with dimensions from one (popular models) to six (He in elliptically polarized fields), as well as multi-channel models (photo-electron spectra for molecules) are treated within the same framework: the degrees of freedom map into a tree hierarchy, inducing a tree-structure for vectors and operators, and producing transparent and efficient code by recursion.

Basis functions are arranged in a tensor-tree with a variety of basis sets, finite-elements, FE-DVR, and grids that can be combined on any number of coordinate axes, with multiple basis sets on the same axis. Discretization and operators can all be input-controlled. For example,

```
#define BOX 20
Axis: name,nCoeff,lower,upper,funcs,order
Phi,5,,,expIm
Eta,3,-1,1, assocLegendre{Phi}
Rn,60, 0,BOX,polynomial,15
Rn,20, BOX,Infty,polExp[0.5]
```

with Eta for $\cos\theta$ defines polar coordinates. The bases $\exp(im\phi)$ and $P_l^{|m|}(\cos\theta)$ combine to the spherical harmonics up to $l = 2$ and the r-coordinate uses 60/15=4 FE-DVR elements on $[0, 20]$ with 15 polynomials on each and 20 exponentially damped polynomials at the end. An example for operator specification is

```
Operator: hamiltonian
0.5(<d_1_d><1>+<1><d_1_d>+<Q*Q><1>+<1><Q*Q>)
```

for $(\overleftarrow{\partial}_x \overrightarrow{\partial}_x + \overleftarrow{\partial}_y \overrightarrow{\partial}_y + x^2 + y^2)/2$. Many standard operators are pre-defined for various coordinate systems.

2.4.4 Software. The code is open source hosted on a Gitlab repository [32]. It is written mostly in C++ and linked with some Fortran-based libraries. Optionally, functionality can be extended by linking FFTW and GSL. Standard compilers are gcc and Intel, ports to Windows and Mac's Clang have been successful but are not actively supported. Compilation is through CMake, and Doxygen documentation is available. Tutorials and further materials are available in the git and from the tRecX website [31].

2.5 XChem

XChem [40, 41] is a solution for an all-electron ab initio calculation of the electronic continuum of molecular systems. XChem combines the tools of quantum chemistry (as implemented, e.g., in MOLCAS [44]) and scattering theory to accurately account for electron correlation in the single-ionization continuum of atoms [41, 43], small and medium-size molecules [42]. At its core lies a close coupling expansion combined with the use of a hybrid Gaussian and B-Spline basis set [40].

This approach yields the scattering states of the molecular system via the eigenstates of the close coupling matrix (CCM). While useful in their own right, the full potential lies in using the close coupling matrix as a starting point for time dependent calculations. In doing so, one may explicitly model the interaction of molecules with ultrashort (attosecond) pulses. The large band widths of such pulses lead to the coherent excitation of multiple ionization channels, whose coupling (accurately described in XChem) gives rise to complex phenomena.

An attractive feature of XChem is that the architecture of the basis functions (Fig. 2) and the use of MOLCAS allow one to describe the electronic continuum of medium-size molecules at the same level of theory as multi-reference CI methods do for the ground and the lowest excited states of such molecules. At present the largest systems treated have of the order of ten atoms.

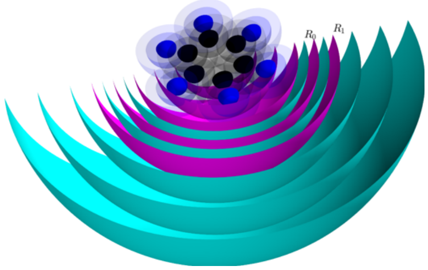


Figure 2: Illustration of the XChem basis architecture in benzene. Cyan: B-splines, mangenta: Gaussians at the center of mass of the molecule, blue and black: Gaussians at the atomic sites not overlapping with B-splines.

2.5.1 What can XChem do? XChem can compute:

- The CCM for a user-defined set of ionization channels (each defined as an ionized molecular state coupled to electrons

of given angular momenta) and including short range states relevant to the problem at hand.
- Scattering states and scattering phases by asymptotic fitting to the analytical solution.
- Total and partial photoionization cross sections within perturbation theory.
- Lifetime and character of resonances embedded in the molecular continuum, either via analysis of the cross section or via inclusion of a complex absorbing potential in the CCM yielding complex eigenenergies.
- The electron dynamics during and after ionization, caused by and probed with ultrashort laser pulses, by solving the time dependent Schrödinger equation using the CCM.
- The angular distribution of photo electrons (in progress).

2.5.2 What can XChem be used for? XChem is a valuable tool for:

- The theoretical study of ultrafast processes in in Attosecond pump-probe experiments, Attosecond Transient Absorption Spectroscopy (ATAS) and Reconstruction of Attosecond Beatings by Interference of Two-Photon Transitions (RABBIT).
- The investigation of photoionization of complex molecules close to threshold, where electron correlation effects are crucial to describe the photoionization cross sections.
- The study of ionization processes intrinsically dependent on electron correlation, like autoionization and Auger decay.
- The computation of potential energy surfaces for the investigation of molecular dynamics during and after fast photoionization events.

2.5.3 Who is using XChem?

- Researchers in (computational) quantum chemistry or molecular physics interested in studying electron dynamics in the ionization continuum of molecules (e.g., photoionization, charge migration, etc).
- Laboratories investigating ultrafast phenomena in many-electron atoms, small and medium-size molecular systems.

2.6 Other Interesting Software

We are also standing up the rather old Many Electron Structure Applications (MESA) code that was developed at Los Alamos in the 1980's and modified to compute electron scattering and photoionization cross sections from polyatomic molecules using the Complex Kohn Method [45]. While this code is old, and in need of significant modernization, it was built to compute electron scattering and photoionization transition matrix elements for general polyatomic molecules, a capability not present in other codes and of interest to many users.

There is also an effort underway to incorporate MOLSCAT [46], a heavy particle collision code for vibrational-rotational scattering in molecules. Table 1 summarizes some of the applications deployed.

3 AMP SCIENCE GATEWAY DEPLOYMENT AND APPLICATION INTEGRATION

The AMP Gateway is deployed using the Apache Airavata Framework. [4] It relies on the Science Gateway Platform as a Service (SciGaP) (https://scigap.org/) hosting services [3] at Indiana University. The SciGaP platform provides gateway services using the multi-tenanted Apache Airavata middleware. The Airavata core enables features such as managing user identity, accounts, authorization provisioning, and the ability to access XSEDE and other high performance computational resources such as XSEDE's Stampede2, Comet and Bridges. These resources are transparently integrated into the AMP gateway and the batch queues are used for scheduling the execution of models using applications and user defined parameters.

3.1 User Accounts, Authentication and Authorization

The AMP gateway user accounts can be created by users by providing a userid and setting a password along with providing email for verification process. In addition to this an automated process using the user institutional login via CI-Logon [47] is also provided which avoids the email verification step. The gateway administrator controls the access to the resources and needs to approve the user for gateway resource access. The users get a "gateway pending role" when they register. The gateway middleware provides authentication and authorization services through the Keycloak [2] identity management system supported by SciGaP. Once the gateway administrator provides approval, the role of the new user will be changed to "gateway-user" which enables access to gateway resources and applications. The gateway additionally provides "admin-read-only" and "admin roles" with their associated permissions for a group to share the admin responsibilities and reporting purposes. A "gateway-user" then, can use gateway services such as creating, monitoring, sharing and cloning experiments (computational simulations). The users can also add their own compute resource allocations using the functions available under "User Settings". The "admin" users have the authorization to control metadata for accessing XSEDE through the gateway "community login", register and deploy applications and their (user) interfaces, manage users, and monitor and access all user experiments. These privileges enable the admin to efficiently troubleshoot any issues relating to the user services. The "admin-read-only" users can view all "admin" related information but not modify any of the settings.

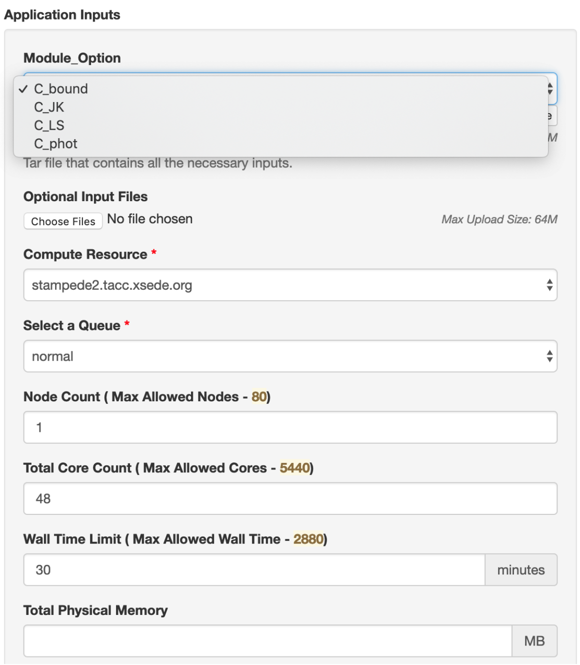
3.2 AMP Application Deployment and User Interface Creation

The AMP gateway started with four specific applications described above: BSR, XChem, CCC and tRecX suites. These applications were compiled and tested on a number of the XSEDE resources. Each of them requires a different set of libraries and in the case of XChem integration with other Open source software such as OpenMolcas [12]. They were independently tested by the collaborating XSEDE ECSS consultant when deployed by the developers or deployed by the consultant to establish the required environments and tweak make/cmake data. The Application deployment in the

Code	Application	Method	Parallel or Serial	Access	Restrictions
BSR	Electron scattering, photoionization, structure	R-matrix/B-spline	S, MPI	BSR (serial)	atoms, atomic ions
CCC	Electron scattering, Positron scattering Photoionization	Close coupling, Fredholm equations in momentum space	OpenMP and MPI		Quasi one- and two-electron atoms and ions
UKRMol(+)	Electron scattering, photoionization	R-matrix, Close coupling Gaussian and B-spline basis	OpenMP and MPI	Public (Zenodo)	Molecules and clusters ($\leq$ 30 atoms)
tRecX	Strong field photo-emission spectra, rates	TDSE (tSurff, haCC): grids, CI-states, FE-DVR, bases	MPI	Public (Git)	Small molecules, two-electron atoms
XChem	Scattering states Photoionization	Close coupling Configuration Interaction Hybrid Gaussian and B-spline basis	OpenMP	Upon request	Small and medium-size molecules
MESA	Electronic structure, scattering, photoionization	SCF, MCSCF, CI, Complex Kohn	S	By request	Small to medium molecules

Table 1: Some characteristics of the software suites deployed in the AMP Gateway.

AMP Gateway consists of defining the application as a "module", an "interface" is defined to user interaction with the application and a resource specific "deployment" description to fully define it on the gateway. The user interfaces are tailored to each application and enable the users to provide input parameters using files or variables that either can be sent as arguments to the application or a wrapping script. Currently, the interface generator, the PHP Airavata client API, provides ways to define file URIs, variables (strings, real/integer/Booleans) for inputs and standard Error/Outputs and file and variables for outputs. Multiple choices for different application sub-modules can be provided to specify and invoke a specific component of the application(for example, Bound/Photonics/LS/JK under software BSR). This can be defined using a simple comma delimited set and can be used in a wrapper to pass it as arguments or specified as input parameters for the application as depicted in figure 3. The interfaces deployed for the four applications will be further enhanced with additional details for input variables and ingesting file sets as archives and whole folders in due course. The job submission interface then enables users to define HPC resource details such as the system/machine ID that is automatically sorted from the list obtained from the deployment description, queue/partition and allocation registered by the administrator or a user, and job specific details such as the number of nodes and processors, time and memory specifications. In additions these experiment specification the Apache Airavata framework provides a sharing mechanism for jobs/experiments and projects (which are collections of experiments) to be shared with collaborators and can be set during the experiment creation (or at any time after). This allows collaborative job submission, monitoring and analysis of the results.



Application Inputs

Module_Option

✓ C_bound
C_JK
C_LS
C_phot

Tar file that contains all the necessary inputs.

Optional Input Files

Choose Files No file chosen
Max Upload Size: 64M

Compute Resource *
stampede2.tacc.xsede.org

Select a Queue *
normal

Node Count (Max Allowed Nodes - 80)
1

Total Core Count (Max Allowed Cores - 5440)
48

Wall Time Limit (Max Allowed Wall Time - 2880)
30 minutes

Total Physical Memory
MB

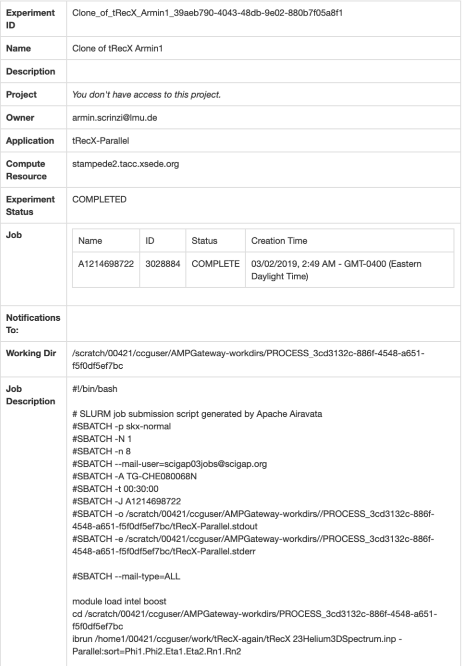
Figure 3: Input interface for BSR software to select a specific module for execution and provide inputs as a tarball and set the resource requirements

3.3 Monitoring Job Progress

During experiment creation users can provide their email to receive messages at job start and end supplied by the scheduler. Additionally, once the experiment is accepted and launched, an "Experiment Summary" interface is launched and automatically refreshed periodically in order to show the status of the job submitted into the XSEDE resources. Currently the status of the job in the scheduler is reflected in the summary interface shown in figure 4. Gateway users can monitor experiments owned by them or shared with

Schneider, Barry I.; Bartschat, Klaus; Zatsarinny, Oleg; Bray, Igor; Martin, Fernando; Scrinzi, Armin; Pamidighantam, Sudhakar; Tennyson, Jonathan; Gorfinkiel, Jimena; Klinker, Markus. "A Science Gateway for Atomic and Molecular Physics." Paper presented at Practice and Experience in Advanced Research Computing (PEARC), Chicago, IL, US. July 28, 2019 - August 01, 2019.

them by other gateway users. Gateway administrators can monitor all gateway experiments using the "Experiment Statistics" page available in the "Admin Dashboard". The monitoring information is provided by a log processing system that extracts the relevant experiment task level execution logs from the gateway middleware and presents it in the gateway monitoring interfaces as depicted in figure 5.



<table>
<tr><td><b>Experiment ID</b></td><td>Clone_of_tRecX_Armin1_39aeb790-4043-48db-9e02-880b7f05a8f1</td></tr>
<tr><td><b>Name</b></td><td>Clone of tRecX Armin1</td></tr>
<tr><td><b>Description</b></td><td></td></tr>
<tr><td><b>Project</b></td><td><i>You don't have access to this project.</i></td></tr>
<tr><td><b>Owner</b></td><td>armin.scrinzi@lmu.de</td></tr>
<tr><td><b>Application</b></td><td>tRecX-Parallel</td></tr>
<tr><td><b>Compute Resource</b></td><td>stampede2.tacc.xsede.org</td></tr>
<tr><td><b>Experiment Status</b></td><td>COMPLETED</td></tr>
<tr><td><b>Job</b></td><td><table><tr><td>Name</td><td>ID</td><td>Status</td><td>Creation Time</td></tr><tr><td>A1214698722</td><td>3028884</td><td>COMPLETE</td><td>03/02/2019, 2:49 AM - GMT-0400 (Eastern Daylight Time)</td></tr></table></td></tr>
<tr><td><b>Notifications To:</b></td><td></td></tr>
<tr><td><b>Working Dir</b></td><td>/scratch/00421/ccguser/AMPGateway-workdirs/PROCESS_3cd3132c-886f-4548-a651-f5f0df5ef7bc</td></tr>
<tr><td><b>Job Description</b></td><td><pre>
#!/bin/bash

SLURM job submission script generated by Apache Airavata
#SBATCH -p skx-normal
#SBATCH -N 1
#SBATCH -n 8
#SBATCH --mail-user=scigap03jobs@scigap.org
#SBATCH -A TG-CHE080068N
#SBATCH -t 00:30:00
#SBATCH -J A1214698722
#SBATCH -o /scratch/00421/ccguser/AMPGateway-workdirs//PROCESS_3cd3132c-886f-4548-a651-f5f0df5ef7bc/tRecX-Parallel.stdout
#SBATCH -e /scratch/00421/ccguser/AMPGateway-workdirs//PROCESS_3cd3132c-886f-4548-a651-f5f0df5ef7bc/tRecX-Parallel.stderr

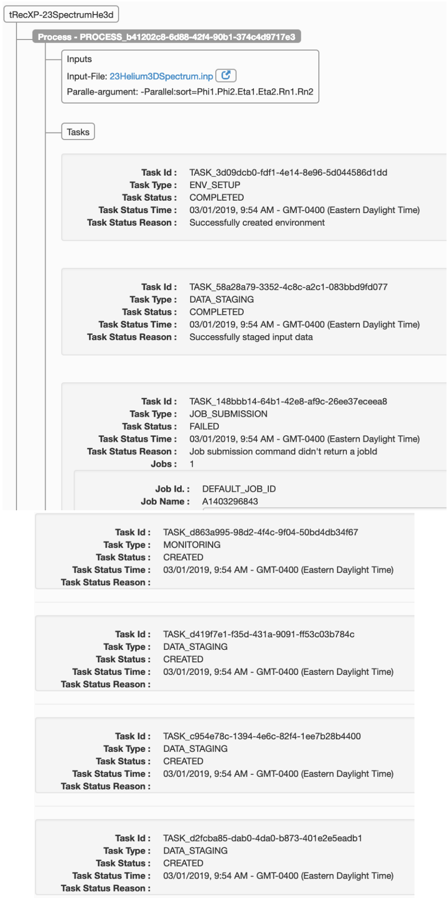
#SBATCH --mail-type=ALL

module load intel boost
cd /scratch/00421/ccguser/AMPGateway-workdirs/PROCESS_3cd3132c-886f-4548-a651-f5f0df5ef7bc
ibrun /home1/00421/ccguser/work/tRecX-again/tRecX 23Helium3DSpectrum.inp -Parallel:sort=Phi1.Phi2.Eta1.Eta2.Rn1.Rn2
</pre></td></tr>
</table>

Figure 4: Experiment summary for an job with the batch script created by Airavata middleware

3.4 Gateway Admin Dashboard

The "Admin Dashboard" is the workspace for the gateway administrator(s) within the gateway. All the admin features discussed above are available through the Admin Dashboard. Apart from what has been discussed, the admin dashboard provides a notification feature and also a way to managing gateway preferences when it comes to individual compute resource and storage resource connectivity and also helps managing credentials for secure compute resource communications. Gateway administrators can create notifications for gateway users and share them with the users with set begin and expiration times. The gateway admin can set and define preferences for the usage of each compute resource with specifications such as community or shared login name, job scratch location for the job data staging and execution, preferred job submission and file transfer protocols, and allocation project number for charging the run time. Similarly, the admin dashboard is used to generate an SSH credential token and key to be used for authentication and authorization at the compute resource and storage resource communications. The admin dashboard also provides job scripts created by the Airavata middleware, the actual path of the job directory on the remote HPC system and detailed task level logs for a specific job to the administrator to check health of the job workflow or troubleshoot in case it is needed.



Input-File: 23Helium3DSpectrum.inp
Paralle-argument: -Parallel:sort=Phi1.Phi2.Eta1.Eta2.Rn1.Rn2

```
           Task Id :  TASK_3d09dcb0-fdf1-4e14-8e96-5d044586d1dd
         Task Type :  ENV_SETUP
       Task Status :  COMPLETED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :  Successfully created environment

           Task Id :  TASK_58a28a79-3352-4c8c-a2c1-083bbd9fd077
         Task Type :  DATA_STAGING
       Task Status :  COMPLETED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :  Successfully staged input data

           Task Id :  TASK_148bbb14-64b1-42e8-af9c-26ee37eceea8
         Task Type :  JOB_SUBMISSION
       Task Status :  FAILED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :  Job submission command didn't return a jobId
              Jobs :  1
              Job Id. :  DEFAULT_JOB_ID
             Job Name :  A1403296843

           Task Id :  TASK_d863a995-98d2-4f4c-9f04-50bd4db34f67
         Task Type :  MONITORING
       Task Status :  CREATED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :

           Task Id :  TASK_d419f7e1-f35d-431a-9091-ff53c03b784c
         Task Type :  DATA_STAGING
       Task Status :  CREATED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :

           Task Id :  TASK_c954e78c-1394-4e6c-82f4-1ee7b28b4400
         Task Type :  DATA_STAGING
       Task Status :  CREATED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :

           Task Id :  TASK_d2fcba85-dab0-4da0-b873-401e2e5eadb1
         Task Type :  DATA_STAGING
       Task Status :  CREATED
  Task Status Time :  03/01/2019, 9:54 AM - GMT-0400 (Eastern Daylight Time)
Task Status Reason :
```

Figure 5: Admin interface showing task level log for an experiment

4 COMMUNITY BUILDING

A proposal has been written to the MOLecular Software Sciences Institute(MOLSSI) [48] to host a series of workshops designed to ;

1. (1) Promote our ideas to a larger more diverse group of scientists than the ITAMP workshop participants both in A&M physics as well as other related fields to help us solidify our ideas.
2. (2) Initially conduct a three day workshop, most likely hosted by NIST sometime this fall.
   - We envision inviting about 30 participants consisting of both A&M scientists and quantum chemists.

- We have requested support from the MOLSSI for the participants and have already been informed that we can expect $15K to support our efforts.
- The NSF Computational Physics program has also promised $10K in support.

(3) The workshop will consist of a number of general sessions discussing the codes and what they can and cannot do. In addition, there will be hands on sessions on how to use the codes on the Science Gateway.

(4) Develop a road-map for other workshops focusing more on a specific code or codes for specialists.

5 ACKNOWLEDGEMENTS

BIS acknowledges the Mathematical Software group of the Applied and Computational Mathematics Division at NIST for supporting this work. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) Stampede2, Comet, and Bridges resources, and an ECSS collaboration grant the through the allocation TG-PHY180023, which is supported by National Science Foundation grant number ACI-1548562. Additional support is provided by the Science Gateways Community Institute, NSF Award 1547611 [49]. The SciGaP.org platform and Apache Airavata are supported by NSF Award 1339774. KB and OZ acknowledge NSF support through grant Nos. PHY-1520970, PHY-1803844, and OAC-1834740. FM and MK acknowledge support by the ERC Proof-of-Concept Grant No. 780284-Imaging XCHEM within the Horizon 2020 Framework Program and by the Spanish MINECO Grant No. FIS2016-77889R. JT and JDG acknowledge support through the UK-AMOR high end computing consortium under EPSRC grant EP/R029342/1.

REFERENCES

[1] J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott and N. Wilkins-Diehr 2014 Xsede: Accelerating scientific discovery, CiSE, 6 pp 62âĂŞ74

[2] M. A. Christie, A. Bhandar, S. Nakandala, S. Marru., E. Abeysinghe, S. Pamidighantam, and M. E. Pierce, 2017. Using Keycloak for Gateway Authentication and Authorization.

[3] M. E. Pierce, S. Marru, E. Abeysinghe, S. Pamidighantam, M. Christie, and D. Wannipurage, 2018, July. Supporting Science Gateways Using Apache Airavata and SciGaP Services. ACM. Science Gateways as a Platform. Proceedings of the Practice and Experience on Advanced Research Computing (p. 99) (accessed 1/30/2019).

[4] M. E. Pierce, S. Marru, E. Abeysinghe, S. Pamidighantam, M. Christie, D. Wannipurage, Supporting Science Gateways Using Apache Airavata and SciGaP Services. In Proceedings of the Practice and Experience on Advanced Research Computing, ACM: Pittsburgh, PA, USA, 2018; pp 1-4.

[5] ITAMP the Institute for Theoretical Atomic and Molecular Physics is an NSF supported institute at the Harvard-Smithsonian which has entered its third decade of existence. ITAMP, as the leading center of AMO physics theory in the US, has developed a notable reputation for training, mentoring and sponsoring postdoctoral and visiting fellows in theoretical AMO science, holding timely workshops and recently, a Winter Graduate School. ITAMP facilitates closer interactions between AMO theory and experiment and has catalyzed cross-disciplinary exchanges between AMO and condensed matter physics.

[6] O. Zatsarinny, BSR: B-spline atomic R-matrix codes, Comp. Phys. Commun. **174** (2006) 273

[7] K. A. Berrington, W. B. Eissner and P. H. Norrington, RMATRX1: Belfast atomic R-matrix codes, Comp. Phys. Commun. **92** (1995) 290

[8] N. R. Badnell, http://amdpp.phys.strath.ac.uk/rmatrix/

[9] P. G. Burke, *R-Matrix Theory of Atomic Collisions*, Springer-Verlag (2011)

[10] O. Zatsarinny and K. Bartschat, The B-spline R-matrix method for atomic processes: application to atomic structure, electron collisions and photoionization J. Phys. B **46** (2013) 112001

[11] Time dependent recursive indexing Software tRecx

[12] OpenMolcas is an open source quantum chemistry code based on multiconfiguration metyhods such as CASSCF and CASPT2

[13] XCHEM is a code that is designed to compute the interaction of atoms and molecules with strong, short pulse electromagnetic fields.

[14] I. Bray and A. T. Stelbovics, Adv. Atom. Mol. Phys. **35**, 209 (1995).

[15] I. Bray and A. T. Stelbovics, Phys. Rev. A **46**, 6995 (1992).

[16] I. Bray, Phys. Rev. A **49**, 1066 (1994).

[17] D. V. Fursa and I. Bray, Phys. Rev. A **52**, 1279 (1995).

[18] D. V. Fursa and I. Bray, J. Phys. B: At. Mol. Opt. Phys. **30**, 5895 (1997).

[19] A. S. Kadyrov and I. Bray, Phys. Rev. A **66**, 012710 (2002).

[20] A. S. Kadyrov, C. M. Rawlins, A. T. Stelbovics, I. Bray, and M. Charlton, Phys. Rev. Lett. **114**, 183201 (2015).

[21] A. S. Kadyrov and I. Bray, J. Phys. B: At. Mol. Opt. Phys. **49**, 222002 (2016)
.

[22] Moore L R *et al* 2011 *J. Mod. Optics* **58** 1132

[23] J. M. Carr and P. G. Galiatsatos and J. D. Gorfinkiel, A. G. Harvey, M. A. Lysaght, D. Madden, Z. Mašín, M. Plummer and J. Tennyson, Eur. Phys. J. D, **66**, 58 (2012)

[24] A. F. Al-Refaie and J. Tennyson, Comput. Phys. Commun., **214**, 216 (2017)

[25] J. Tennyson, Phys. Rep., **491**, 29 (2010)

[26] D Darby-Lewis, Z. Mašín and J. Tennyson, J. Phys. B, **50**, 17501 (2017)

[27] D.S. Brambila, A.G. Harvey, K. Houfek, Z. Mašín and O. Smirnova, Phys. Chem. Chem. Phys. (2017) **19** 19673-19682

[28] A. Loupas and J. D. Gorfinkiel, J. Chem. Phys. **150** 064307 (2019)

[29] Z. Mašín, J. Benda, A. G. Harvey, J. D. Gorfinkiel and J. Tennyson, Comput. Phys. Comm., in preparation

[30] J. Tennyson, D. B. Brown, J. Munro, I. Rozum, H. N. Varambhia and N. Vinci, J. Phys. Conf. Ser., **86**, 012001 (2007)

[31] "The tRecX Homepage," ().

[32] "The tRecX git repository," ().

[33] A. Zielinski, V. P. Majety, and A. Scrinzi, Phys. Rev. A **93**, 023406 (2016).

[34] V. P. Majety and A. Scrinzi, Journal of Physics B: Atomic, Molecular and Optical Physics **48**, 245603 (2015).

[35] V. P. Majety and A. Scrinzi, Phys. Rev. Lett. **115**, 103002 (2015).

[36] V. P. Majety and A. Scrinzi, Phys. Rev. A **96**, 053421 (2017).

[37] A. Scrinzi, Phys. Rev. A **81**, 053845 (2010).

[38] A. Scrinzi, New Journal of Physics **14**, 085008 (2012)
.

[39] V. P. Majety, A. Zielinski, and A. Scrinzi, New. J. Phys. **17** (2015), 10.1088/1367-2630/17/6/063002

[40] C. Marante, L. Argenti and F. Martín, Phys. Rev. A **90**, 012506 (2014)

[41] C. Marante, M. Klinker, I. Corral, J. González-Vázquez, L. Argenti, and F. Martín, J. Chem. Theory Comp. **13**, 499 (2017)

[42] M. Klinker, C. Marante , L. Argenti, J. González-Vázquez, and F. Martín, J. Phys. Chem. Lett. **9**, 756 (2018)

[43] C. Marante, M. Klinker, T. Kjellsson, E. Lindroth, J. González-Vázquez, L. Argenti, and F. Martín, Phys. Rev. A. 96, 022507 (2017)

[44] F. Aquilante, J. Autschbach, R. K. Carlson, L. F. Chibotaru, M. G. Delcey, L. De Vico, N. Ferré, L. M. Frutos, L. Gagliardi, M. Garavelli, et al, J. Comp. Chem. **37**, 506 (2016)

[45] T. N. Rescigno, C. W. McCurdy, A. E. Orel and B. H. Lengsfield III, The Complex Kohn Variational Method in, "Computational Methods for Electron-Molecule Collisions", 1 (1995)

[46] MOLSCAT, is a program for computing non-reactive quantum scattering calculations involving atomic and molecular particles. See Jeremy M. Hutson, C. Ruth Le Sueur

[47] CILogon is an open source, standards-based service providing the NSF research community with credentials for secure access to cyberinfrastructure (CI)

[48] Molecular Sciences Software Institute is an organization located in Blacksburg, VA which serves as a nexus for science, education, and cooperation serving the worldwide community of computational molecular scientists – a broad field including of biomolecular simulation, quantum chemistry, and materials science.

[49] Pierce, M.E., Marru, S., Gunathilake, L., Wijeratne, D.K., Singh, R Wimalasena, C., Ratnayaka, S. and Pamidighantam, S., 2015. Apache Airavata: design and directions of a science gateway framework. Concurrency and Computation: Practice and Experience, 27(16), pp. 4282-4291.

Performance Study of a GAA-GAA Coexistence Scheme in the CBRS Band

Weichao Gao and Anirudha Sahoo
National Institute of Standards and Technology
Gaithersburg, Maryland, U.S.A.
Email: {weichao.gao anirudha.sahoo}@nist.gov

***Abstract*—The General Authorized Access (GAA) users in the Citizens Broadband Radio Service (CBRS) band are the lowest priority users who not only have to make sure that they do not cause harmful interference to the higher tier users but also must cooperate with each other to minimize potential interference among themselves. Thus, efficient GAA coexistence scheme is essential for operation of GAA users and to obtain high spectrum utilization. Towards this goal, the Wireless Innovation Forum (WInnForum) has recommended three schemes to facilitate coexistence among the GAA users. To the best of our knowledge, there is no performance study on any of these schemes available in the public domain. In this paper, we study performance of one of these schemes (called Approach 1). We choose two actual locations in the USA around which our study is conducted using actual terrain and land cover data of the continental USA. We evaluate performance of the scheme at different deployment densities, using different propagation models and with different mix of CBRS devices (CBSDs) at those two locations. We provide some interesting insights into the bandwidth allocation process and performance of Approach 1 in terms of mutual interference.**

I. Introduction

The Federal Communications Commission (FCC) in the USA has published the rules for commercial use of the spectrum in the 3.5 GHz band known as Citizens Broadband Radio Service (CBRS) band on a sharing basis [1]. The CBRS band has a three tiered access model. Current incumbents will operate in the highest tier followed by the Priority Access License (PAL) users in the middle tier and the General Authorized Access (GAA) users in the lowest tier. The incumbents must be protected from harmful interference caused by tier-2 (PAL) and tier-3 (GAA) users. PAL users should be protected from interference from GAA users. However, a GAA user cannot expect interference protection from higher tier users as well as from other GAA users in the same tier. Access to the spectrum in this band is managed by Spectrum Access Systems (SASs). As per the rule 47 C.F.R. § 96.35 in [1], GAA users must cooperate with each other to minimize the potential interference and to increase spectrum utilization. In the first phase of deployment in the CBRS band, there will be no PAL users. Hence, only GAA users will share the spectrum with the incumbents. Thus, GAA-GAA coexistence is very criticial to the success of this band. The Wireless Innovation Forum (WInnForum), which is involved in developing standards for operation of systems in the CBRS band has published Technical Reports recommending different schemes to faciliate effective GAA-GAA coexistence that should minimize mutual interference and increase spectrum utlization. The WInnForum has recommended three different schemes for GAA-GAA coexistence in three different Technical Reports [2]–[4]. The design and architecture of these schemes are largely based on discussions and experience of various members of the WInnForum. To the best of our knowledge, there is no performance study on any of these schemes available in the public domain. In this paper, we take up one of those schemes, named *Approach 1*, proposed in [2] and study its performance in different configurations. In the CBRS band, there can be two types of CBRS devices (CBSDs). Category A (CatA) CBSDs transmit at lower power than Category B (CatB) and are typically installed indoors. CatB CBSDs are deployed outdoors. We study the effect of propagation model, deployment density and different population of CatA and CatB CBSDs on the performance of the GAA-GAA coexistence. It is envisioned that operators will group their CBSDs into, what are called, Coexistence Groups (CxGs). The CxGs will be responsible for managing interference among their respective CBSDs. Hence, a SAS will only be responsible for allocating bandwidth to the CxGs.

The main contributions of this work are as follows. The WInnforum does not define any performance metric to evaluate its proposed schemes. We have proposed a few performance metrics, which will be useful for operators and SAS administrators to evaluate the schemes as well as to compare different schemes. To the best of our knowledge, there is no such study on the GAA-GAA coexistence schemes available in the public domain. Consequently, our work should provide insight into the performance of the scheme (Approach 1) proposed in [2] in terms of various deployment parameters and propagation models. We use actual deployment location data and use the WInnForum reference implementation of propagation models [5] which uses actual terrain and land cover data of the continental USA. Hence, our simulation results should be close to what one would expect in practice. As explained later in the paper, in one of our experiments, we deviate from the WInnForum scheme and show how more bandwidth (compared to WInnForum scheme) can be allocated at the cost of higher interference. Results from this experiment suggest that a better scheme can be designed to provide more bandwidth to the CBSDs if they agree to tolerate higher

Gao, Weichao; Sahoo, Anirudha. "Performance Study of a GAA-GAA Coexistence Scheme in the CBRS Band." Paper presented at IEEE Dynamic Spectrum Access Networks (DySPAN) Conference, Newark, NJ, US. November 11, 2019 - November 14, 2019.

TABLE I: List of Acronyms

CBRS	Citizens Broadband Radio Service
PAL	Priority Access License
GAA	General Authorized Access
SAS	Spectrum Access System
CBSD	CBRS device
CxG	Coexistence Group
CIG	CBSD Interference Graph
EW	Edge Weight
ET	Edge Threshold
BW	Bandwitdh
IM	Interference Metric
VB	Virginia Beach
SD	San Diego
ITM	Irregular Terrain Model
SIRG	Signal to Interference Ratio at a Grid
AIPA	Average Interference Power per unit Area
AIPCCG	Average Interference Power per CBSD per Channel per grid

interference up to a certain threshold.

II. Related Work

Coexistence issues in different wireless bands have been studied in the past. Coexistence challenges for heterogeneous cognitive networks in the TV white space have been discussed in [6]. In this study, coexistence between the secondary users and the incumbents as well as among the secondary users is discussed. Coexistence among secondary users which are heterogeneous in their air interface and MAC protocol is also considered. Coexistence of LTE-licensed assisted access (LTE-LAA) and WiFi in the 5 GHz band has been studied in [7]. Coexistence of LTE-LAA and WiFi in the TV white space has been proposed in [8], [9]. Some of the solutions proposed in the literature are to modify LTE MAC protocol to improve coexistence performance. The above coexistence scenarios are addressed with specific air interface or MAC protocol in mind. However, the GAA-GAA coexistence schemes in the CBRS band proposed by the WInnForum do not assume any particular air interface or MAC protocol. As mentioned earlier, the WInnForum has proposed three approaches to solve the GAA-GAA coexistence problem. Approach 1 [2] treats bandwidth as the only resource and hence, allocates bandwidth to the CBSDs such that interfering CBSDs are assigned different channels to the extent possible. It does not manipulate transmit power of the CBSDs for coexistence purpose. If the deployment is too dense and hence, assigning different channels to interfering CBSDs is not possible, then this scheme allows some CBSDs to be assigned the same channel even if they may interfere with each other. Approach 2 [3] deals with bandwidth and transmit power together and treats them as two types of resources. In dense deployment scenarios, if there are not enough channels to allocate different channels to interfering CBSDs, then less transmit power is allocated to a pair of interfering CBSDs so that intereference between them is mitigated and hence, can be allocated the same channel. Approach 3 [4] tries to maximize the amount of bandwidth allocated to individual CxGs by using a recursive algorithm to a cluster of CBSDs. It first identifies the CBSDs which belong to a CxG and are only connected to (i.e., interfere with) CBSDs which belong to the same CxG. These CBSDs are refered to as *cluster of size 1*. These clusters can be allocated $100\,\%$ of the available bandwidth. Next is to identify CBSDs belonging to cluster of size 2. CBSDs in these clusters belong to one of two CxGs. In this case, $50\,\%$ of available bandwidth is allocated to CBSDs belonging to one CxG and the other $50\,\%$ is allocated to CBSDs belonging to the other CxG. This algorithm is then applied recursively until all CBSDs are covered. A study of impact of propagation models on GAA-GAA coexistence and deployment density is presented in [10].

III. Overview of WInnForum Scheme (Approach 1)

The WInnForum has proposed three different schemes as solutions to GAA-GAA coexistence. In this section, we present salient parts of one of these schemes, named Approach 1 [2], which we have used in our study.

A. CBSD Interference Graph

For the purpose of GAA-GAA coexistence, a CBSD Interference Graph (CIG) is constructed in a deployment area. The vertices in the CIG are the CBSDs. An edge is placed between two CBSDs if either one or both of the CBSDs experience interference from the other CBSD above a given threshold. Edge Weight (EW) between all pairs of CBSDs is computed to determine if an edge should exist between the pair. If the computed EW is above a set Edge Threshold (ET), then an edge is established between the two CBSDs.

1) Edge Weight Calculation: For Edge Weight (EW) calculation, an Interference Metric (IM) between two CBSDs is first computed. IM is a measure of mutual interference between two CBSDs. Depending on the deployment scenario, IM may be computed in *area coordination* or in *point coordination* mode. For example, when CBSDs are deployed as LTE e-NodeB, then it needs to have a coverage area which should be protected from interference. Hence, in this case, IM in area coordination mode should be computed. On the other hand, when two CBSDs are deployed for Fixed wireless service, one CBSD is deployed as the Base Transceiver Station (BTS) and the other is deplyed as a Customer Premise Equipment (CPE) CBSD. They communicate in point-to-point mode and hence, interference at those CBSDs needs to be limited. The point coordination mode is appropriate in this case. In this study, we are intereseted in CBSD deployment for LTE coverage and hence, focus on area coordination mode.

In area coordination mode, for a pair of CBSDs, say CBSD-1 and CBSD-2, coverage area of each CBSD is computed. Coverage area of a CBSD, for a given transmit power, is the area around the CBSD such that the received signal strength at any point inside the area is above a set threshold. The WInnforum scheme specifies that this threshold should not be less than -96 dB relative to 1 mW (dBm)/10 MHz. The fraction of coverage area of CBSD-1 that overlaps with the coverage area of CBSD-2 is taken as CBSD-1's interference metric IM_1. Similarly, interference metric IM_2 of CBSD-2 is the overlap area expressed as a fraction of its coverage area. Then the EW between CBSD-1 and CBSD-2 is the maximum

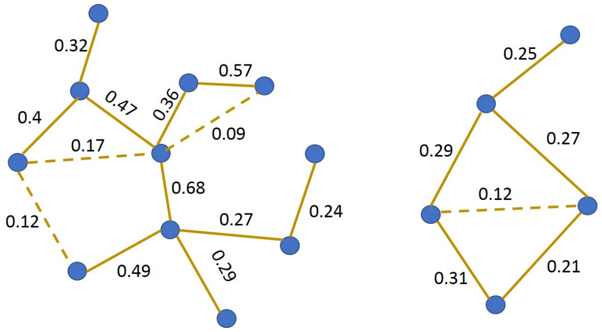


Fig. 1: An Example CBSD deployment with Edge Weights

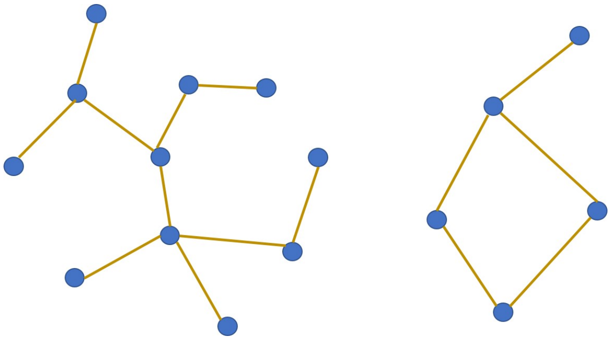


Fig. 2: Example CBSD Interference Graph when ET=0.2

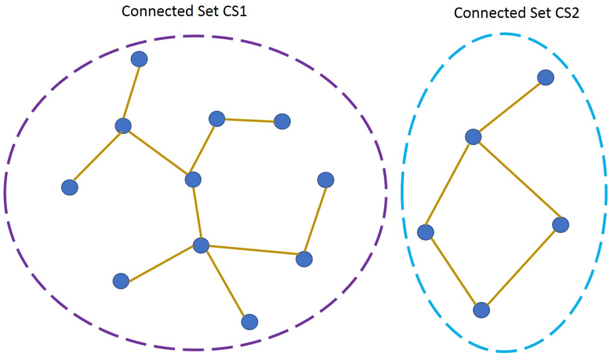


Fig. 3: Example Connected Sets

of IM_1 and IM_2. Note that EW takes a value between 0 to 1. For a given edge threshold (ET), an edge is established between CBSD-1 and CBSD-2 only if the EW is greater than the ET. This procedure is followed for every pair of CBSDs to obtain the CBSD interference graph.

2) Connected Set: Once the CBSD inferference graph is constructed, the next step is to generate *connected set(s)* off of it. A CBSD interference graph may contain one or more connected sets. Any two CBSDs in a connected set are connected directly through an edge or indirectly through other CBSDs in the interference graph. No CBSD within a connected set is connected directly or indirectly to any CBSD outside of the connected set [2].

Fig. 1 shows an example of CBSD Interference Graph when the ET is set to 0.2. In the figure, there is a solid edge between two CBSD if their coverage areas overlap and the EW between them is greater than or equal to the ET. A dashed edge indicates that the coverage areas of the two CBSDs overlap, but the EW is less than the ET. No edge between two CBSDs implies that the coverage area of the two CBSDs do not overlap. After applying edge threshold and removing the dashed edges, we get the CBSD interference graph as shown in Fig. 2. When the conditions of connected set are applied to this interference graph, we get two connected sets CS1 and CS2 as shown in Fig. 3.

3) Coexistence Groups: It is envisioned that operators in this band will create Coexistence Groups (CxGs) to faciliate GAA-GAA coexistence. A CxG consists of a group of CBSDs which will coordinate their own interference within the group. Thus, a SAS is only responsible for the allocation of bandwidth at the CxG level. The operator (or a CxG manager) of a CxG will take the bandwidth allocated to it and assign it to individual CBSDs within the CxG as per its interference management policy. As a result, a connected set will consist of one or more CxGs, i,e., CxGs are subgraphs in a connected set. The CBSDs which do not belong to any CxG are grouped together to form a common CxG (sort of a virtual CxG).

4) Graph Coloring of Connected Sets: The WInnForum scheme proposes a graph coloring approach [11] to allocate GAA bandwidth. The graph coloring starts at the CxG sub graph level. Graph coloring of a CxG involves computing its *chromatic number*. Chromatic number of a CxG is the minimum number of colors required to color the nodes of the CxG such that no two nodes having an edge between them are assigned the same color. Once chromatic number of each CxG inside a connected set is computed, then the *total chromatic number* of the connected set is computed by summing up the chromatic numbers of the CxGs belonging to the connected set. The bandwidth allocation to the CxGs is done as per the following procedure [2].

Let B be the total bandwidth available for the GAA users. Let C_i be the chromatic number of CxG_i. If there are M CxGs in the connected set, then the total chromatic number of the connected set is $C = \sum_{i=1}^{M} C_i$ and the bandwidth allocated to CxG_i is given by

$$BW_i \quad = \quad B \cdot \frac{C_i}{C} \tag{1}$$

Note that the bandwidth allocated to a CBSD is B/C. It is understood that for useful operation, a CBSD should get at least 10 MHz bandwidth. Consequently, if $B/C < 10$ MHz then the ET needs to be increased which will eliminate some edges from the connected set and hence, bring down the value of C. Then the bandwidth allocation process is repeated again. This procedure is repeated until $B/C \geq 10$ MHz.

TABLE II: CBSD Parameters

Area Type	Antenna Height [m] (Above Ground Level)		EIRP [dBm/10MHz]	
	Cat A	Cat B	Cat A	Cat B
Dense Urban	50%: 3 to 15 25%: 18 to 30 25%: 33 to 60	6 to 30	26	40 to 47
Urban	50%: 3 50%: 6 to 18	6 to 30	26	40 to 47
Suburban	70%: 3 30%: 6 to 12	6 to 100	26	47
Rural	80%: 3 20%: 6	6 to 100	26	47

TABLE III: Ratio of CBSD categories deployed in different Areas

Area Type	Cat A	Cat B
Dense Urban	90 %	10 %
Urban	90 %	10 %
Suburban	90 %	10 %
Rural	95 %	5 %

IV. Experiment and Results

A. Deployment Model

We consider a deployment area of $5\,km \times 5\,km$ in size around Virginia Beach (VB) in the east coast (the center at latitude 36.872227 and longitude -76.023389) and around San Diego (SD) in the west coast (the center at latitude 32.723588 and longitude -117.145319) of the USA.

We chose these two cities because the terrain around these two cities are quite different. The terrain around Virginia Beach is somewhat flat, whereas it is hilly around San Diego. Propagation loss is a function of the terrain profile between transmitter and receiver. Hence, the two chosen cities have quite different propagation characteristics. The coverage area of CBSDs are clipped by the above square deployment area. The deployment area is discretized by dividing it into equal sized grids of size $50\,m \times 50\,m$. CBSDs are uniformly placed around this deployment area as per the deployment density used for a given experiment. The parameters of the CBSDs used in our experiments are shown in Table II. All the CBSDs are assumed to have omnidirectional antennae.

In this study, we have assumed that each CBSD is a singleton CBSD and hence, all the CBSDs in the deployment area form one CxG. Since the FCC rule allows up to 70 MHz (out of total of 150 MHz) for PAL users, we assume that the rest 80 MHz is available for GAA users. We have used -96 dBm/10 MHz as the receive power threshold while computing coverage area of a CBSD.

B. Deployment Configurations

We ran our experiments in two deployment configurations as follows.

- *Config A*: In this configuration, all the deployed CBSDs are chosen to be Category A.

TABLE IV: ITM Parameters

Parameter	Value
Polarization	1 (Vertical)
Dielectric constant	25 (good ground)
Conductivity (S/m)	0.02 (good ground)
Mode of Variability (MDVAR)	13 (broadcast point-to-point)
Surface Refractivity (N-units)	ITU-R P.452
Radio Climate	ITU-R P.617
Confidence/Reliability Var. (%)	50/50

- *Config B*: In this configuration, we used a mix of Category A and Category B CBSDs as per Table III.

For each of the above configurations, we ran experiments with different deployment densities and propagation models at the two chosen locations (SD and VB). All the Category A CBSDs in our experiements are considered indoors, whereas all the Category B CBSDs are deployed outdoors.

C. Performance Metrics

The WInnForum does not suggest any performance metrics for evaluating the GAA-GAA coexistence scheme. In this section, we describe the performance metrics used in our evaluations.

- *Signal to Interference Ratio at a Grid (SIRG)*: The signal power at a grid within the coverage area of one or more CBSDs operating on the same channel is the highest received power at the grid from one of these CBSD transmitters. The received power at that grid from all other CBSDs operating on the same channel is considered as interference power. So, the SIRG (for a given channel) is the ratio of signal power to the aggregate Interference power expressed in dB.
- *Average Interference Power per unit Area (AIPA)*: This metric captures the average interference experienced by a receiver while it is inside the coverage area of a CBSD. If there are N_g grids inside the coverage area of a CBSD and I_i is the interference power (in dBm) received at the grid i over a channel assigned to the CBSD, then the AIPA (in dBm) of the CBSD, on that channel is given by

$$AIPA \quad = \quad 10\log_{10}\left(\frac{\sum_{i=1}^{N_g} 10^{I_i/10}}{N_g}\right) \qquad (2)$$

- *Average Interference Power per CBSD per Channel per grid (AIPCCG)*: The AIPCCG is defined as the average interference power (in dBm) per CBSD per channel per grid. Let I_i^j be the interference power (in dBm) received at a grid i on channel j. Let N_g, N_c and N_d be the number of grids, channels and CBSDs in the deployment area respectively. Then AIPCCG is given by

$$AIPCCG \quad = \quad 10\log_{10}\left(\frac{\sum_{j=1}^{N_c}\sum_{i=1}^{N_g} 10^{I_i^j/10}}{N_g \cdot N_c \cdot N_d}\right) \quad (3)$$

D. Propagation Models

We have evaluated performance of the GAA-GAA coexistence scheme using two different propagation models: the Irregular Terrain Model (ITM) (in point to point mode) [12] and the Hybrid model as described in the Requirement R2-SGN-04 in [13]. The ITM model, also known as the Longley-Rice model, is a propagation model based on electromagnetic theory, terrain features and radio measurements. The parameters used in the ITM propagation model are given in Table IV. The Hybrid propagation model is a model proposed by the WInnForum and is a hybrid between the ITM and the extended Hata (eHata) model. The eHata model [14] is an extension of the Hata model [15], which is essentially an empirical model based on a series of land-mobile measurements made by Okumura [16] over varied terrain. While the eHata model accounts for clutter loss, the ITM model does not consider clutter loss. The Hybrid propagation model primarily sets its loss equal to the larger of the ITM loss and the eHata loss in urban and suburban area. In the rural area, the propagation loss using the Hybrid model is equal to the loss using the ITM model. Thus, in general, propagation loss using the Hybrid model is higher than or equal to the ITM model.

E. Bandwidth Allocation

In this section, we analyze bandwidth allocated to the CBSDs using the WInnForum GAA-GAA coexistence scheme (Approach 1). The experiments were run for all combinations of locations (San Diego and Virginia Beach), propagation models (ITM and Hybrid), and deployment densities of 3, 10, 30 and 50 $CBSDs/km^2$ for both the CBSD deployment configurations.

We first analyze the BW allocation in San Diego. Fig. 4 and Fig. 5 show the cumulative distribution function (CDF) of BW allocation for different deployment densities and propagation models for Config A and Config B respectively. For both Config A and B, for the ITM model as deployment density increases we generally see better BW allocation to the CBSDs (more CBSDs get more BW). This is counter intuitive. However, in these cases, as deployment density increases, the interference graph becomes more connected leading to a higher chromatic number. If the chromatic number is too high then the BW allocated to the CBSDs goes below 10 MHz and hence, the algorithm increases the ET which results in a lower chromatic number. In some cases, some edges in a connected set may be eliminated so as to create multiple connected sets due to this. Each connected set has the entire available bandwidth at its disposal for allocation to its CBSDs. Hence, the increase in the ET increases the chance of getting more BW for each individual CBSD. An extreme case is when the algorithm has to raise the ET to 1.0 for some of the connected sets (as is the case for the ITM model at density 50 in Config B). In this case, each CBSD becomes a single-CBSD connected set and is assigned the entire available bandwidth. As we will discuss later, this improvement in BW allocation comes at the cost of incurring higher interference. However, in the case of the Hybrid propagation model, as the density increases, there is no clear trend in BW allocation. The way the BW allocation algorithm is designed, when deployment density increases, as described above, the system parameters such as the ET, the number of connected sets, the chromatic number in connected sets change. With so many parameteric changes in the system, it is hard to predict a trend in the BW allocation when the deployment density increases. Comparing BW allocation using the Hybrid vs the ITM propagation model, again, it is hard to conclude which model produces better performance. In Config A, the Hybrid model produces better performance whereas in Config B, the ITM gives better performance. In general, the Hybrid propagation model produces equal or more loss than the ITM. So, one can generally assume to get better BW allocation than the ITM. However, sometimes the ITM model can result in better BW allocation (see in case of Config B) (at the cost of higher interference as we will see later). Comparing performance between Config A and B, we see that if the Hybrid model is used, Config A gives better performance, but if the ITM model is used, then there is no clear winner. Again, this is because there are many system parameters that change when the deployment density increases.

At Virginia Beach (Fig. 6 and Fig. 7), the BW allocation does not vary that much when the deployment density increases, especially for Config B (in which the interference graph is more connected due to higher transmission power of CatB CBSDs). Because of the flat terrain, propagation loss is less compared to SD. Hence, CBSDs quite far away are connected to each other. As a result, degree of vertices is high for low deployment density. As the deployment density increases, the degree of vertices does not increase significantly. Thus, the chromatic number of connected sets, and consequently BW allocation does not increase significantly. Note that chromatic number of a graph is less than or equal to the (maximum vertex degree +1). When the deployment density increases, we see that the BW allocation remains the same or becomes better for both propagation models and for both Config A and B. The reason is same as we discussed before for SD. For other cases, there is no clear trend.

When we compare BW allocation between SD and VB, generally SD has an equal or a better BW allocation for both configurations. This can be attributed to the hilly terrain around SD which leads to more propagation loss which, in turn, leads to less dense connectivity in the connected sets and hence, results in a lower chromatic number.

For both SD and VB, it is hard to draw a trend in the BW allocation across two configurations. As mentioned before, when the deployment density increases, there are multiple system parameters that change. Hence, when using the WInnforum coexistence scheme it is hard to determine which configuration is better in terms of the BW allocation.

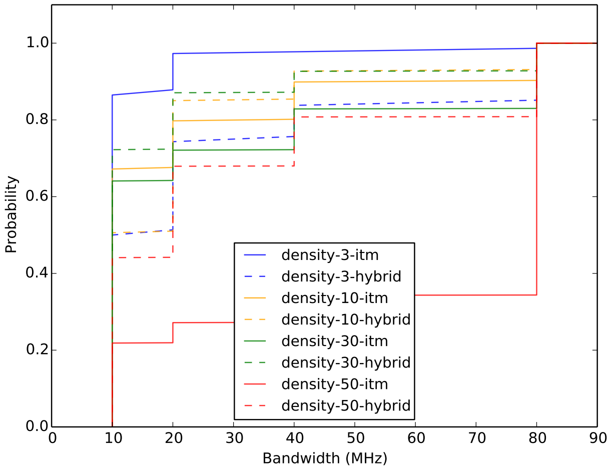


Fig. 4: CDF of Bandwidth Allocated for Config A (San Diego)

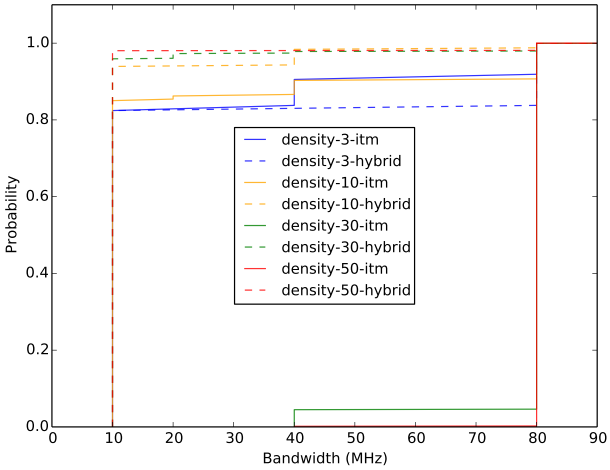


Fig. 5: CDF of Bandwidth Allocated for Config B (San Diego)

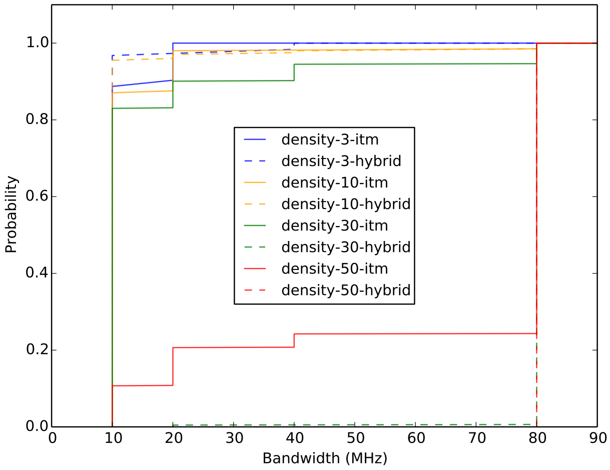


Fig. 6: CDF of Bandwidth Allocated for Config A (Virginia Beach)

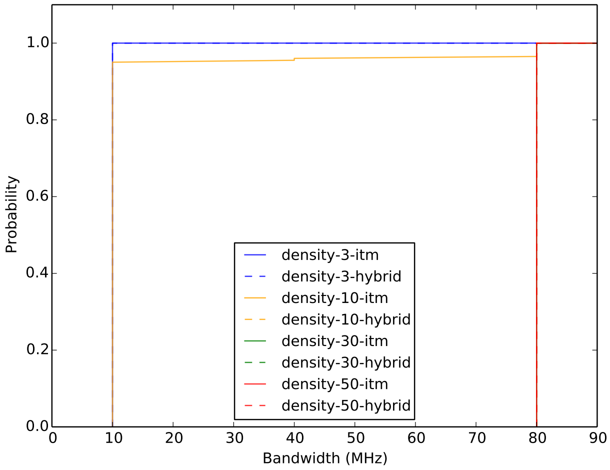


Fig. 7: CDF of Bandwidth Allocated for Config B (Virginia Beach)

F. Performance in terms of AIPA

Fig. 8 and Fig. 9 show the CDF of the AIPA with different propagation models and deployment densities in SD in Config A and Config B respectively for the channel with the worst interference. The corresponding figures for VB are Fig. 10 and Fig. 11. For a given propagation model and a given configuration, as the deployment density increases the AIPA becomes worse for both SD anf VB location. This is quite intuitive. When the deployment density increases, there is more interference due to transmission from higher number of CBSDs which causes the AIPA to increase. Another factor that contributes to the AIPA increase is when the scheme has to increase the ET to allocate a minimum of 10 MHz BW to the CBSDs. As explained in the BW allocation for SD, the ITM model allocates more BW when the deployment density increases, but this is achieved at the cost of increasing the ET which leads to a higher AIPA. We see this increase in the AIPA for the ITM model in SD in Fig. 8 and Fig. 9. In SD, the Hybrid propagation model results in better AIPA than the ITM model in both the configurations. But in VB, the ITM does better than the Hybrid in both the configurations. This is because

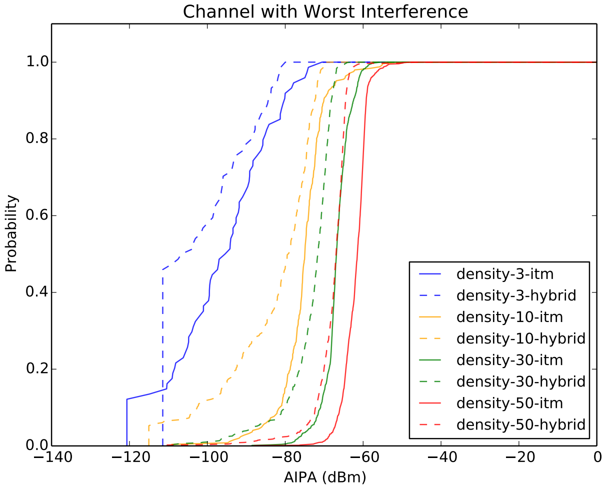


Fig. 8: CDF of AIPA for Config A (San Diego)

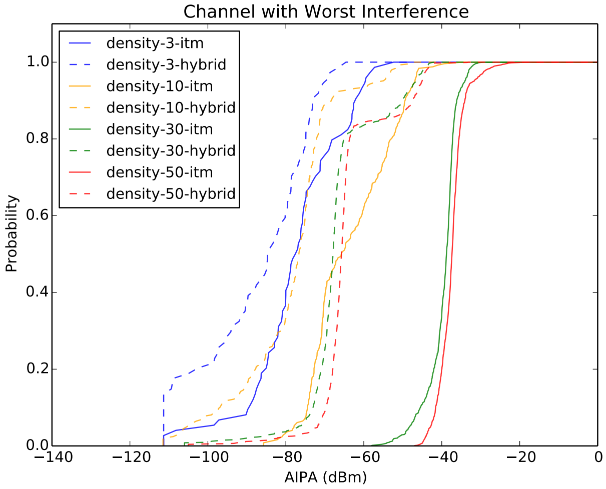


Fig. 9: CDF of AIPA for Config B (San Diego)

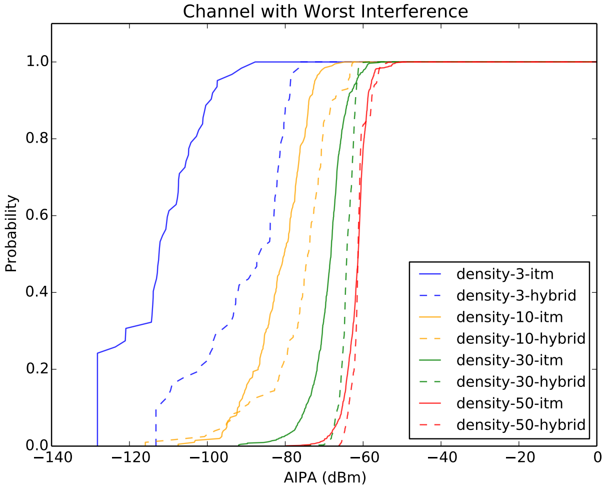


Fig. 10: CDF of AIPA for Config A (Virginia Beach)

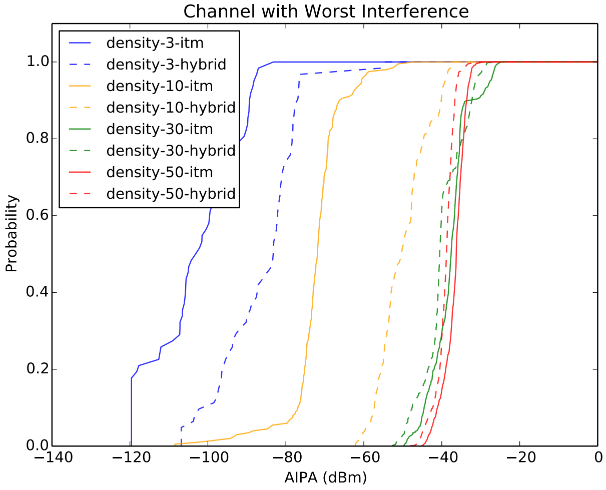


Fig. 11: CDF of AIPA for Config B (Virginia Beach)

of the way the Hybrid model is defined. When the Hybrid model is used in urban and suburban areas, the propagation loss takes on the value provided by the eHata (since its loss is generally more than the ITM in such areas) whereas in rural areas the propagation loss is equal to that provided by the ITM (as per Requirement R2-SGN-04 in [13]). In SD, the majority grids are in urban or suburban area, so that the propagation loss is determined by the eHata model in most cases when the Hybrid model is used, which leads to higher loss. Thus, the AIPA in SD is better for the Hybrid model than when the ITM model is used. In contrast VB has a large rural area. Thus, when the Hybrid model is used, the propagation loss in VB is mostly equal to that calculated by the ITM model. As a result, one would expect the AIPA performance of the ITM and the Hybrid model to be very close to each other. However, as per the implementation of the Hybrid model by the WinnForum (see R2-SGN-04 in [13] and [5]), antenna height of a CBSD cannot be less than 20 m. Due to this requirement, for the CBSDs having height less than 20 m typically its coverage using the ITM would be lower than that using the Hybrid model. Lower coverage area leads to lower interference. Since VB is dominated by rural grids, the ITM provides better AIPA performance than the Hybrid.

For both the locations, the AIPA performance is better for Config A for both the propagation models and for all deployment densities. This is intuitve, since having Cat B

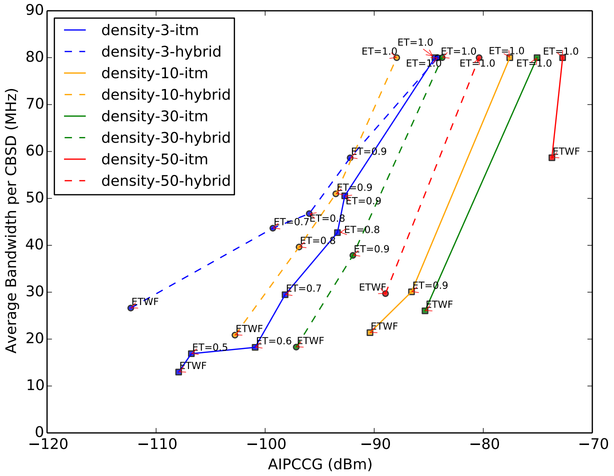


Fig. 12: BW Allocation vs AIPCCG at Different ET in Config A (San Diego)

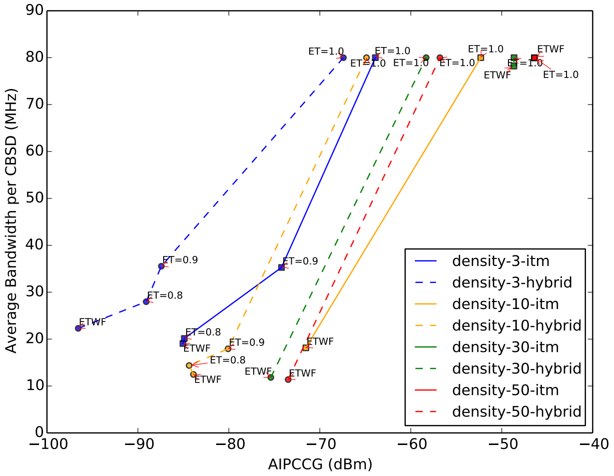


Fig. 13: BW Allocation vs AIPCCG at Different ET in Config B (San Diego)

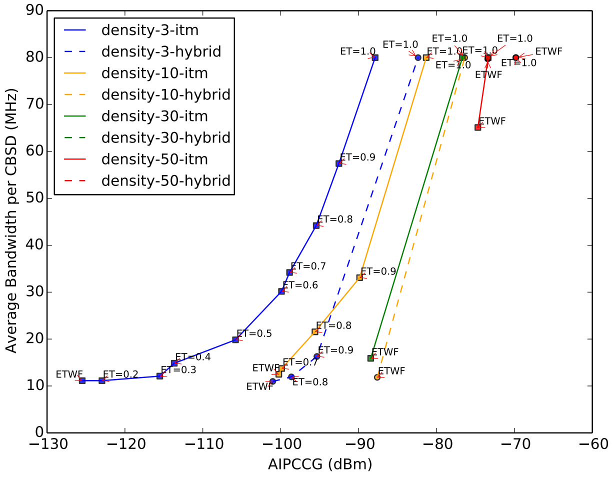


Fig. 14: BW Allocation vs AIPCCG at Different ET in Config A (Virginia Beach)

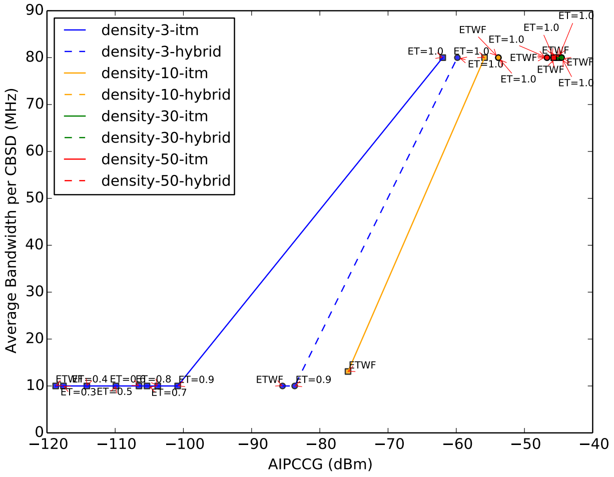


Fig. 15: BW Allocation vs AIPCCG at Different ET in Config B (Virginia Beach)

CBSDs (which have higher transmit power), creates more interference.

G. Performance of BW Allocation vs AIPCCG

For this performance measurement, we deviate from the scheme proposed by the WInnForum. In this experiment, we want to observe the effect of allocating higher BW at the cost of higher interference when we go beyond the ET at which the proposed WInnForum scheme would stop. Note that the proposed WInnForum scheme stops increasing the ET of a connected set once the CBSDs in the connected set get at least 10 MHz bandwidth. Figures 12, 13, 14 and 15 show how increase in the ET results in more average BW allocation per CBSD at the cost of interference for the locations SD (Config A and B) and VB (Config A and B) respectively. Note that in this experiment, the CBSDs are allocated actual BW computed for a given ET, i.e., the final BW allocation is not rounded down to multiples of 10 MHz. The points marked as ETWF (WInnForum ET) represents the operating point of the WInnForum Scheme in terms of average BW and AIPCCG. Note that, in general, there will not be a single ET value at this operating point since there could be multiple connected

Gao, Weichao; Sahoo, Anirudha. "Performance Study of a GAA-GAA Coexistence Scheme in the CBRS Band." Paper presented at IEEE Dynamic Spectrum Access Networks (DySPAN) Conference, Newark, NJ, US. November 11, 2019 - November 14, 2019.

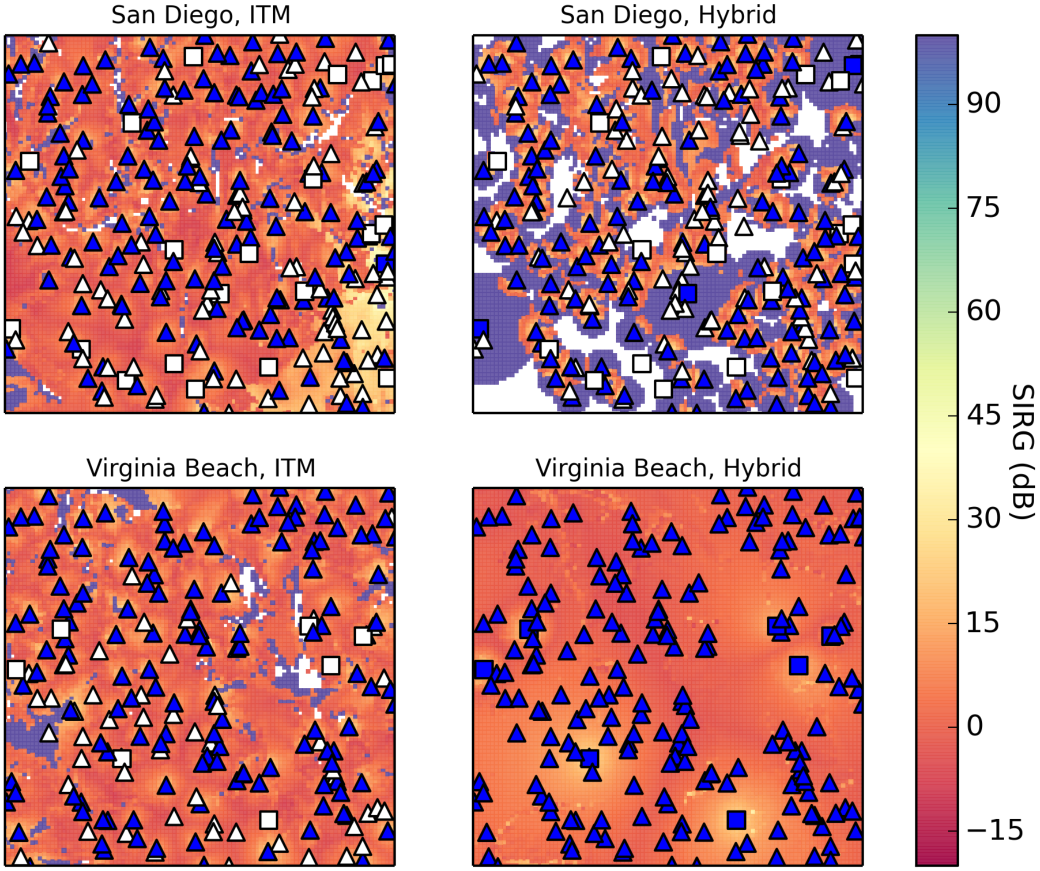


Fig. 16: Heatmap of SIRG of Channel 1, Deployment Density 10, Config B

sets each with its own ET. Hence, we do not provide an ET value at this operating point in the graphs. For a given deployment density, we then continue to increase the ET beyond the corresponding ETWF. The interference metric in this experiment is AIPCCG and its computation is explained in Section IV-C. At both SD and VB location, as expected, when a higher BW is allocated to the CBSDs, the AIPCCG also goes up for all combinations of configurations, propagation models and deployment densities. Also, as the deployment density increases, to get the same BW allocation, the ET needs to be higher and the corresponding AAIPC is also higher. For a given propagation model and a given deployment density, as the ET increases, the CBSDs get more BW at the cost of higher AIPCCG. In SD, the Hybrid propagation model produces a better result than the ITM model for all deployment densities and for both the configurations, i.e., for a given allocated BW, the AIPCCG is lower for the Hybrid model than the ITM model. But in VB, the ITM model produces a better BW allocation than the Hybrid model. This reversal of performance between the two propagation models at the two locations is due to the same reason as explained in the performance in terms of the AIPA.

H. Performance in terms of SIRG

Figure 16 shows the SIRG heatmap in SD and VB for both propagation models in Config B for the most crowded channel (Channel 1). The CBSDs in blue have been allocated channel 1 whereas the CBSDs in white operate on some other channel(s). The color coded scale (in dB) is provided to the right. For both the ITM and the Hybrid propagation models, SD shows better SIRG performance over VB, although the difference is more prominent with the Hybrid model. With the hilly terrain around SD, the Hybrid propagation loss is more in SD than in VB. Hence, there is less interference which leads to a higher SIRG. At SD, using the Hybrid model gives a much better SIRG performance over the ITM. Since SD has lot of urban and suburban areas, the Hybrid model incurs more loss and hence, leads to less interference. At VB, there is no marked difference between the ITM and the Hybrid propagation in terms of SIRG. VB has a vast rural area in which the ITM and the Hybrid model produce almost the same loss. Hence, the SIRG performance using those two models at VB has no significant difference. We have the SIRG performance for other deployment densities and configurations, but we are not able to present them here due to space limitation.

V. Conclusion and Future Work

In this paper, we studied the performance of the proposed WInnForum GAA-GAA coexistence scheme, called Approach 1. Our study looked at the effect of propagation model, deployment density and different mix of CatA and CatB CBSDs on the performance of GAA-GAA coexistence. Our study found that the way WinnForum Approach 1 is designed, performance of the BW allocation is hard to predict. There are multiple system parameters at play while allocating BW (e.g., ET, number of CSs and chromatic number of each CS), which are inter-dependent, making the prediction hard. In terms of the AIPA performance, Config A performs better than Config B. The AIPA performance in SD is better than in VB for the Hybrid model, whereas the ITM model performs better in VB. The way the GAA-GAA coexistence scheme is designed, it is possible to get a better BW allocation at higher deployment densities at the cost of incurring higher interference. The SIRG performance is generally better at locations having hilly terrian (e.g., SD) than at locations having flat land. At a given location, the Hybrid model will generally have a better SIRG performance than the ITM model.

From our study of performance of average BW vs AIPCCG, we feel a better scheme could be to have a target threshold for the SIRG in a deployment area and then allocate the maximum possible BW to CBSDs such that the SIRG does not go below the threshold. We intend to study this scheme further and compare its performance with WInnForum's Approach 1. We would like to analyze the performance of the WInnForum scheme when the CBSD deployment has multiple CxGs to investigate the effect of having CxGs on BW allocation and on interference. We are in the early stage of implementing WInnForum's Approach 3 proposed in [4]. We would like to compare the performance of that scheme with Approach 1.

References

[1] "Citizens Broadband Radio Service," 47 C.F.R. § 96, 2019.

[2] "Operations for Citizens Broadband Radio Service (CBRS); GAA Spectrum Coordination - Approach 1," Document WINNF-TR-2003, Version V1.0.0, May 2019. [Online]. Available: https://winnf.memberclicks.net/assets/work_products/Recommendations/WINNF-TR-2003-V1.0.0%20GAA%20Spectrum%20Coordination-Approach%201.pdf

[3] "Operations for Citizens Broadband Radio Service (CBRS); GAA Spectrum Coordination - Approach 2," Document WINNF-TR-2004, Version V1.0.0, May 2019. [Online]. Available: https://winnf.memberclicks.net/assets/work_products/Recommendations/WINNF-TR-2004-V1.0.0%20GAA%20Spectrum%20Coordination-Approach%202.pdf

[4] "Operations for Citizens Broadband Radio Service (CBRS); GAA Spectrum Coordination - Approach 3," Document WINNF-TR-2005, Version V1.0.0, May 2019. [Online]. Available: https://winnf.memberclicks.net/assets/work_products/Recommendations/WINNF-TR-2005-V1.0.0%20GAA%20Spectrum%20Coordination%20-%20Approach%203.pdf

[5] "Reference models for SAS testing." [Online]. Available: https://github.com/Wireless-Innovation-Forum/Spectrum-Access-System/tree/master/src/harness/reference_models

[6] C. Ghosh, S. Roy, and D. Cavalcanti, "Coexistence challenges for heterogeneous cognitive wireless networks in TV white spaces," *IEEE Wireless Communications*, vol. 18, no. 4, pp. 22–31, 2011.

[7] B. Chen, J. Chen, Y. Gao, and J. Zhang, "Coexistence of LTE-LAA and Wi-Fi on 5 GHz with corresponding deployment scenarios: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 7–32, 2016.

[8] F. M. Abinader, E. P. Almeida, F. S. Chaves, A. M. Cavalcante, R. D. Vieira, R. C. Paiva, A. M. Sobrinho, S. Choudhury, E. Tuomaala, K. Doppler *et al.*, "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Communications Magazine*, vol. 52, no. 11, pp. 54–61, 2014.

[9] A. M. Cavalcante, E. Almeida, R. D. Vieira, S. Choudhury, E. Tuomaala, K. Doppler, F. Chaves, R. C. Paiva, and F. Abinader, "Performance evaluation of LTE and Wi-Fi coexistence in unlicensed bands," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*. IEEE, 2013, pp. 1–6.

[10] Y. Hsuan, "Impacts of Propagation Models on CBRS GAA Coexistence and Deployment Density," in *Invited Presentation, WInnComm*, 2018. [Online]. Available: https://winnf.memberclicks.net/assets/Proceedings/2018/Invited%20Hsuan.pdf

[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.

[12] "Irregular Terrain Model (ITM) (Longley-Rice) (20 MHz–20 GHz)." [Online]. Available: https://www.its.bldrdoc.gov/resources/radio-propagation-software/itm/itm.aspx

[13] "Requirements for Commercial Operation in the U.S. 3550–3700 MHz Citizens Broadband Radio Service Band," Wireless Innovation Forum Document WINNF-TS-0112, Version V5.0, Mar. 2018. [Online]. Available: https://workspace.winnforum.org/higherlogic/ws/public/document?document_id=4743&wg_abbrev=SSC

[14] E. Drocella, J. Richards, R. Sole, F. Najmy, A. Lundy, and P. McKenna, "3.5 GHz Exclusion Zone Analyses and Methodology," National Telecommunications and Information Administration, Technical Report TR 15-517, Mar. 2016. [Online]. Available: http://www.its.bldrdoc.gov/publications/2805.aspx

[15] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE transactions on Vehicular Technology*, vol. 29, no. 3, pp. 317–325, 1980.

[16] Y. Okumura, "Field strength and its variability in VHF and UHF land-mobile radio service," *Rev. Electr. Commun. Lab.*, vol. 16, pp. 825–873, 1968.

Simulation Testbed for Railway Infrastructure Security and Resilience Evaluation

Himanshu Neema
Xenofon Koutsoukos
himanshu.neema@vanderbilt.edu
xenofon.koutsoukos@vanderbilt.edu
Institute for Software Integrated Systems
Vanderbilt University
Nashville, TN 37235

Bradley Potteiger
brad.potteiger@jhuapl.edu
Johns Hopkins Applied Physics Laboratory
Laurel, MD 20723

CheeYee Tang
Keith Stouffer
cheeyee.tang@nist.gov
keith.stouffer@nist.gov
National Institute of Standards and Technology
Gaithersburg, MD 20899

ABSTRACT

The last decade has seen an influx of digital connectivity, operation automation, and remote sensing and control mechanisms in the railway domain. The management of the railway operations through the use of distributed sensors and controllers and with programmable and remotely controllable railway signals and switches has led to gains in system efficiency as well as operational flexibility. However, the network connectivity has opened up the railway cyber communication networks to cyber-attacks. These are a class of cyber-physical systems (CPS) with interconnected physical, computational, and communication components. The cyber-attacks on these systems could potentially cascade through these interconnection and result into significant damage. These systems are safety-critical owing to their large-scale monetary and, more importantly, human life safety concerns. Therefore, it is better to incorporate security and resilience requirements right from the design time. In this paper, we describe a domain-specific framework for simulations in the railway domain. The framework allows analyzing the resilience of railway operations in the presence of cyber-attacks. In particular, our simulation framework allows modeling the railway network as well as the railway transportation. It provides an online graphical modeling environment that allows multiple users to collaborate, through a web-based interface, over the same model for the railway infrastructure as well as network attacks. The framework also allows the user to configure and run experiments through the web-interface and also to visualize the key operational metrics from the railway domain as the experiment is running. The framework also supports executing large simulations in the cloud. In addition, it supports hardware-in-the-loop (HIL) simulation for incorporating physical effects and network attacks that can only be realized realistically in the hardware. A detailed case study is provided to demonstrate the framework's capabilities.

CCS CONCEPTS

• **Computer systems organization** → **Embedded and cyber-physical systems**; *Redundancy*; • **Computing methodologies** → **Modeling and simulation**; • **General and reference** → *Cross-computing tools and techniques*; • **Networks** → Network reliability.

KEYWORDS

Railway infrastructure, Modeling and Simulation, Hardware-in-the-loop, Cybersecurity, Resilience

ACM Reference Format:
Himanshu Neema, Xenofon Koutsoukos, Bradley Potteiger, CheeYee Tang, and Keith Stouffer. 2020. Simulation Testbed for Railway Infrastructure Security and Resilience Evaluation. In *HoTSoS '20: 7th Annual Hot Topics in the Science of Security (HoTSoS) Symposium, April 07–08, 2020, Lawrence, KS.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Railway infrastructure is going through a transformation through incorporation of network enabled sensors and actuators that make it possible to control its operations and remotely and in an automated manner. The growth in network connectivity and with programmable and remotely controllable railway signals and switches has made railway operations and management more efficient as well as operationally flexible. At the same time, the open connectivity has made them vulnerable to cyber-attacks. These systems are a class of CPS [14] with interconnected physical, computational, and communication components. These are also safety-critical systems because a large number of transportation and even human life depends on their continual safe operations. As such the railway infrastructure is one of the nationwide critical infrastructure. However, the tight coupling among the communication, computational, and physical components enable cascading failures once one of the component gets compromised and attacked. Thus, even though the integration of vehicle to infrastructure (V2I) and vehicle to vehicle (V2V) enables fine-grained control over the transportation operations, it makes them susceptible for cyber-attacks that can cause significant damage to the systems and can even cause loss of human life. Therefore, analyzing the security and resilience of railway systems is critical for studying the effect of cyber-attacks.

Railway is unique due to the tight integration between legacy standalone devices and modern communication interfaces. As such, many systems which were designed several decades ago, do not take into account the vulnerability space presented by remote

Potteiger, Bradley; Neema, Himanshu; Koutsoukos, Xenofon; Tang, CheeYee; Stouffer, Keith A. "Simulation Testbed for Railway Infrastructure Security and Resilience Evaluation." Paper presented at Hot Topics in the Science of Security (HoTSoS) 2020, Lawrence, KS, US. April 07, 2020 - April 08, 2020.

communication. This fact combined with the rush to market by Internet-of-Things (IoT) manufacturers makes railway susceptible to an increased and diversified attack surface, especially the communication-related vulnerabilities such as memory corruption. As such, the threshold to a successful compromise is significantly lower compared to traditional information technology applications such as servers, websites, and databases. Since in modern railway there is a tight integration between software processes and physical dynamics, vulnerable applications can be exploited for causing physical damage, that could potentially include terrorist attacks by sophisticated adversaries.

One key differentiator between CPS and traditional software applications is the unpredictability in analyzing the impact of cyber-attacks on a live system, particularly the physical actuation of safety-critical components. Additionally, the increased interconnectedness between components increases the potential and impact of attack propagation. As such, it is no longer sufficient to rely on locking down the most critical components, but zero trust architectures must be utilized to take a weakest-link approach to prevent adversaries from gaining entry into critical networks. Finally, the necessity of requiring the stability of physical actuation, makes it important to not only analyze the cyber-attack behavior on the underlying software components, but also the physical operations of systems. As such, in railway applications, a high priority focus should be on system safety, high availability, system and data integrity and predictable operation.

The main problem that this paper addresses is how to leverage simulation and emulation capabilities to create secure and resilient CPS. In addition, we focus on how to provide a systematic methodology for creating cyber-attack experiments to assess their impact, with or without defense mechanisms, on the software and physical dynamics of the system. We also demonstrate integrated evaluation of key metrics and their visualization to provide real-time feedback to security researchers.

In this paper, we describe a domain-specific framework for simulations in the railway domain. The framework allows analyzing the resilience of railway operations in the presence of cyber-attacks. In particular, our simulation framework allows modeling and integrated simulation of the railway cyber communication network and the railway transportation. The framework is developed as an online graphical modeling tool that allows multiple users to collaborate, through a web-based interface, over the same model for the railway infrastructure as well as network attacks. In the framework, we have developed methods to define and calculate key operational metrics to study the impact of cyber-attacks on railway operations. Using this approach, one can design and evaluate different mechanisms for network defense and determine system's key network vulnerabilities. In the following sections we attempt to focus on the following objectives:

- Develop a software platform for rapidly designing and evaluating cyber-attacks for connected railway architectures.
- Design a component-based modeling approach that leverages our modular libraries for deploying cyber-attacks and collecting metrics from domain-specific experiment results in real-time.
- Integrate HIL testbed in order to evaluate the impact of cyber-attacks and physical effects in a hardware environment similar to that in the real-world.
- Create a case study using the Washington, DC Metro railway network model for demonstrating the capabilities of our experiment design platform.

The remainder of the paper is organized as follows. Section 2 provides the rationale for evaluating cybersecurity in the railway domain. Section 3 describes the architecture of our cybersecurity evaluation platform. Section 4 demonstrates the capabilities of our platform utilizing the case study of the Washington, DC Metro system. Section 5 discusses the work related to this research and Section 6 concludes.

2 RATIONALE

With the push to smart city implementations, railway transportation has experienced a significant disruption. New fuel sources, communication protocols, and control systems have rapidly increased the efficiency and safety of these systems, while reducing the operational cost. Trains now comprise several embedded electronic equipment dedicated for both internal operations and remote access purposes such as communicating with local infrastructure and central monitoring stations. With increased communication capabilities, manual mechanisms are no longer necessary for track configuration. Similar to the automobile industry, track control systems such as rail signals and switches are built with autonomous logic to systematically route trains. This advanced control requires vehicle to infrastructure (V2I) communication protocols which are largely being expanded due to the implementation of 5G.

Embedded microcontrollers and V2I communication have definitely disrupted the railway industry in a positive manner. However, this change has also negatively impacted cybersecurity of the system. Railway systems, like in the automotive industry, rely on a significant amount of legacy code, most of which has not been adapted in decades due to the safety and regulation costs. Consequently, the railway systems have a larger number of vulnerabilities compared to the traditional information technology infrastructure, which allows for nation states, terrorist organizations, and hacktivists to compromise these highly critical transportation networks. The presence of legacy code and remote communication capabilities effectively means that physical access is no longer necessary to disrupt train operations. Instead, an adversary can easily compromise train systems even when being several miles away. Furthermore, because safety-critical train networks are not isolated, once an attacker obtains access, they can pivot to safety-critical devices, causing them to behave dangerously that could lead to devastating crashes. Some of the high profile examples of attacks on railway systems include ransomware attacks against San Francisco and Sacramento, and system-wide disruption in train scheduling operations in London [12]. Fortunately, there haven't been any high profile examples of cyber-attacks leading to dangerous crashes. However, translating these results to the Amtrak crash in Philadelphia [25] paints a eye-opening picture of potential consequences of disruption of safety-critical control.

To successfully protect railway networks and systems, it is critical to utilize an objective and scientific methodology to evaluate the most effective cybersecurity defense mechanisms, as well as prioritizing relevant attack surfaces to address. As such, simulation has proved to be an effective technique in the CPS domain to quantitatively evaluate the software effects on the physical dynamics. This ensures that safety and security can be designed into system architectures in a cost-effective manner. Furthermore, for a smooth transition to deployment settings, software emulation in a hardware-in-the-loop environment can increase trust in reliable system performance.

2.1 HIL Testbed

In order to measure the impact of security mechanisms on the performance of the Industrial Control Systems (ICS), the US National Institute of Standards and Technology (NIST) has developed an HIL testbed. NIST has also published a guide for implementation of security in ICS [22]. The testbed utilizes Commercial-Off-The-Shelf (COTS) control hardware as well as several simulation tools for emulating realistic scenarios. For this particular application, the HIL testbed is used for measuring the performance impact of cybersecurity mechanisms on railway operations. The testbed uses realistic ICS hardware as well as an integrated framework called the Cyber Physical Systems Wind Tunnel (CPSWT) [8] [9] [11] [18] that enables integrating large-scale heterogeneous simulations. The HIL setup in the NIST testbed emulates a railroad crossing scenario. It has a COTS Programmable Logic Controller (PLC), a commercial industrial network switch, and two embedded sensors. The simulation is hosted in a virtual machine that has an Ethernet connection to the PLC. The PLC has a Controller Area Network (CAN) interface to communicate with the embedded sensors. The CAN protocol is used widely for ICS in the railroad and automotive industries. Transmission Control Protocol (TCP) and Internet Protocol (IP) are the standard protocols used in the worldwide internet. The HIL testbed simulator communicates with the PLC through a TCP/IP socket for sending commands to the PLC and receiving sensor information. The PLC uses the sensor information to determine the train location and speed at a crossing, then uses its output to control the barrier and warning signal at the crossing. The barrier control and warning signal in the testbed are represented by simple analog outputs from the PLC. When an experiment is performed in the simulator, the hardware is functioning in real-time instead of simulation time, allowing for a more accurate representation of the railway behavior in deployment environments. This setup allows the researcher to evaluate any impact induced by the experiment to the crossing.

In the past, we have focused our work on utilizing the CPSWT simulation integration framework to evaluate the cybersecurity of railway scenarios [9]. CPSWT takes advantage of the IEEE High Level Architecture (HLA) standard [2] to integrate various domain-specific simulators synchronously and analyze integrated simulations with different system configurations and parameters in the context of many different CPS experiments. This work takes a step further by developing a cloud-based experiment manager to rapidly develop and evaluate railway specific attack scenarios. Furthermore, we have made our testbed setup more user friendly by utilizing an open source simulator [21], allowing easier creation of transportation scenarios from scratch. The next sections present the architecture of our testbed and demonstrate its capabilities with the use of a railway case study.

3 SYSTEM ARCHITECTURE

Railway transportation system is one of the nation's critical infrastructure as a large number of people travel by trains as well as a large amount of packages and goods are transported by the system. A failure or attack in one of the components of the system can lead to cascading failures in other parts of the system, which can quickly result in substantial financial and human loss. This is the reason why these are safety-critical systems. Therefore, evaluating these systems for safety, reliability, and security in the presence of cyber attacks is necessary.

However, these systems are highly complex with many different types of components that work together. For example, the trains, its engine, train tracks, track signals and switches are physical components that are key part of these systems. There are also computational components such as sensors and controllers that make the train operations possible. In addition, there are many communication networks and devices that form the cyber communication network topology of railway systems. Moreover, humans are integral part of all of its operations and workflows. Therefore, a holistic evaluation of these systems requires one to integrate simulators of each of these parts of the system so that overall system-of-systems (SoS) level studies can be conducted. Fig. 1 shows the setup of our testbed. As shown, the front-end provides a modeling environment where users can design simulation studies and execute them using web-based plugins that execute the simulations in the backend. Additionally, the hardware devices connected to the testbed can be used for attacks and effects more practically and realistically realizable in the real hardware.

There are three main components in our simulation framework: *Experiment Controller*, *Simulation Backend*, and *Simulation Analyzer*. The key aspects for analyzing railway systems in our simulation platform include: the development of simulation experiments, scientific design for resilience, ensuring security and resilience amidst sophisticated real-world attacks, and the use of operational quantitative metrics for analysis. Thus, various approaches for securing the railway infrastructure can be objectively compared. Fig. 2 shows the three components of our framework and the sub-sections below describe them in detail.

3.1 Experiment Controller

The *Experiment Controller* (EC) serves as the main orchestrator of experiments. It allows for designing experiment scenarios including modeling and deployment of cyber-attacks. For modeling purposes, the EC uses a web-based graphical modeling environment (or WebGME) [26]. WebGME provides a metamodeling framework that allows users to customize the experiment modeling language. In addition, WebGME provides a multi-user modeling environment with change tracking and allows multiple designers to work on the same model using a web-browser from different locations. Moreover, WebGME allows incorporating domain-specific plugins that enable model interpretation, generation of executable artifacts such as code, scripts, and configurations, and also execution of simulations on the compute platforms (include cloud backends). Furthermore,

Potteiger, Bradley; Neema, Himanshu; Koutsoukos, Xenofon; Tang, CheeYee; Stouffer, Keith A. "Simulation Testbed for Railway Infrastructure Security and Resilience Evaluation." Paper presented at Hot Topics in the Science of Security (HotSoS) 2020, Lawrence, KS, US. April 07, 2020 - April 08, 2020.

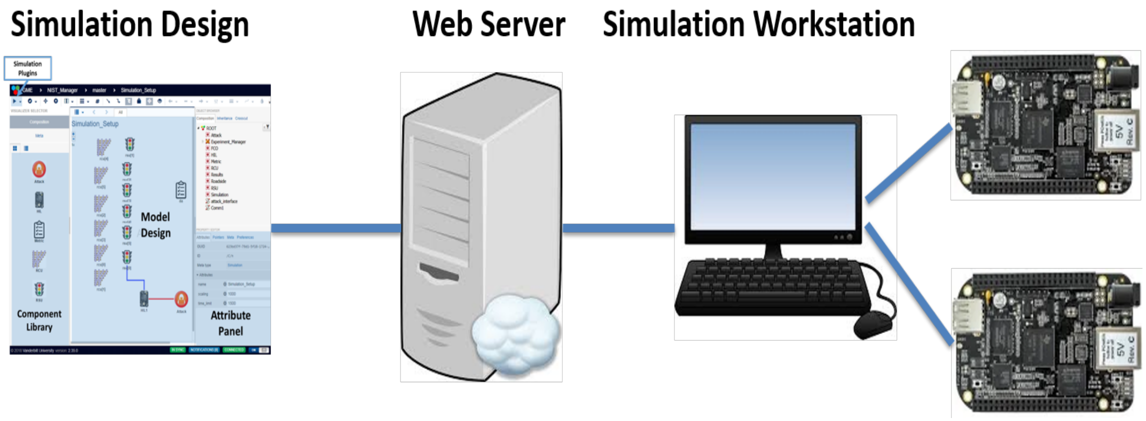


Figure 1: Railway Cybersecurity Evaluation Testbed Setup

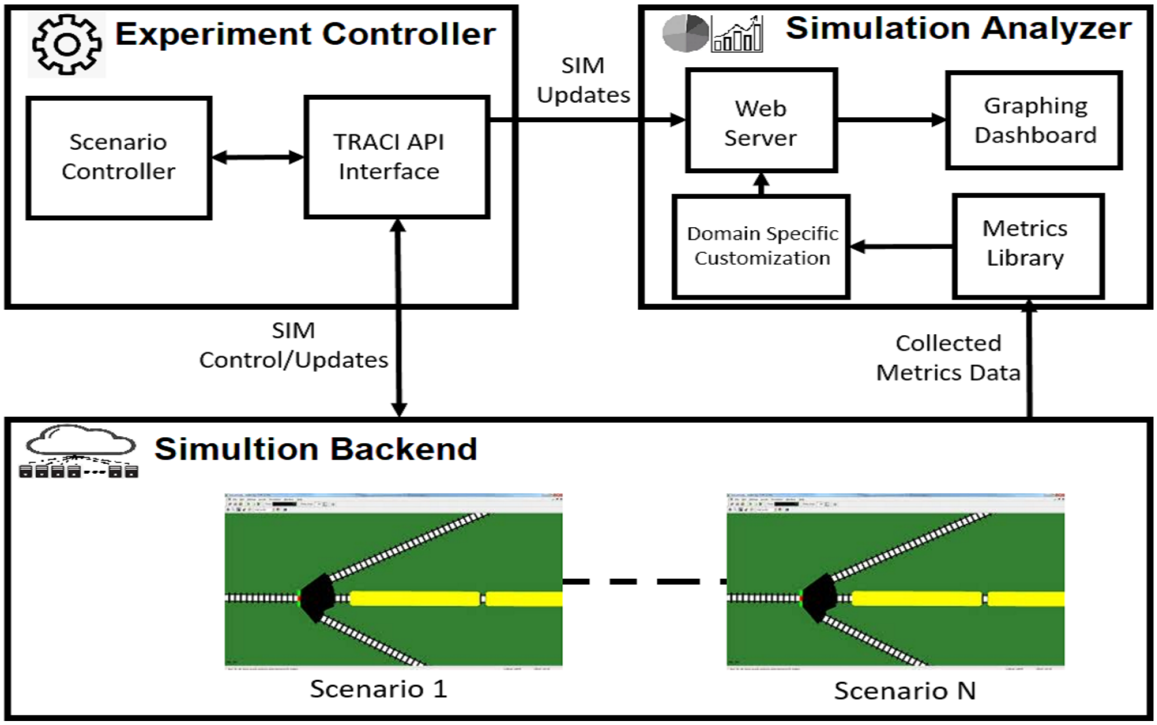


Figure 2: Testbed Architecture

for analyzing the scenarios, the EC monitors the executed simulations, collects the experiment results, and brings it on the front-end for analysis and display tools.

3.2 Simulation Backend

The *Simulation Backend* (SB) is responsible for providing the compute platform and the associated tools and methods for running large-scale experiments. In the SB, we use an open-source simulator, called Veins [21], for modeling and simulating the railway networks and railway transportation as well as the railway infrastructure's cyber communication network. Internally, the Veins simulator is composed of two separate simulators, viz. SUMO [16] for road transportation simulation and OMNeT++ [23] for cyber communication network simulation. With support for simulation experiments to run in the cloud, multiple experiments could be executed at the same time.

3.2.1 Attack Library. An important feature of our framework is that it comes with a reusable set of cyber-attacks that can be configured and utilized in different experiment scenarios. These cyber-attacks are packaged in the form of an attack library. For implementation of the cyber-attacks in the library, we extended the Veins V2X communication module for attack specific source-code and for parameterization of the attacks according to requirements of different experiment scenarios. Currently, we support incorporating denial of service (DOS), integrity, corruption, and delay attacks on railway nodes in simulation, while distributed-DOS (DDOS) attacks can be implemented on the integrated hardware nodes.

3.2.2 HIL Testbed. In order to test for attacks that be deployed only in the hardware, such as DDOS, large amount of network traffic, or attacks exploiting vulnerabilities in the hardware, we integrated our HIL testbed with the simulation framework to provide a rich experimentation environment. The attack library described above already contains such attacks. This environment enables analyzers to develop scenarios where the components can either be simulated in the software or emulated in the physical hardware. For example, in the context of railway simulations, one can design the railway network and transportation in the simulator and emulate railroad signals and actuators in a realistic integrated hardware. Our testbed is comprised of a cluster of multiple Beaglebone Black embedded microcontroller boards running the Ubuntu 16.04 operating systems. Additionally, we support realistic communication protocols by providing capabilities for 100 Mb/s Ethernet and 1 Mb/s CAN bus with the open source ZeroMQ and SocketCAN libraries.

3.3 Simulation Analyzer

The *Simulation Analyzer* (SA) component in our framework enables analysis of experiment results through integrated data capturing and visualization tools. When the users execute the experiment on the Simulation Backend, the experiment artifacts are generated both while the simulation is running and once it has completed. These artifacts are stored in a high-availability database using InfluxDB [3]. The SA has the ability to bring both types of artifacts to the web-based experimentation front-end. In addition, we developed several analysis tools and a presentation dashboard using Grafana [1]. The unique feature of SA is that the analyzed experiment results can be plotted and visualized to the users in real-time. This is accomplished by processing the artifacts as they get generated by the experiments and stored in the database and periodically running the analysis tools on them. Additionally, after experiment conclusion, final results are also plotted in a WebGME visualizer integrated with the user development environment. Moreover, the visualization tools also support comparing different domain-specific experimentation scenarios. For example, one can use these features to perform cyber-gaming experiments in the context of a particular railway scenario by playing various combinations of cyber-attacks against various security mechanisms.

3.3.1 Metrics Library. Domain-specific operational metrics [10] are require computation logic that is specific to the domain as well as the metric. However, many domain-specific aspects can be modularized in a form that becomes reusable for calculating different operational metrics. Our metrics library provides a set of customizable and extensible methods to enable incorporation of newer operational metrics. We implemented this library by integrating a graphical modeling interface with a custom data acquisition module that collects various generic component parameter values from the simulation, and calculates system-level metrics utilizing predefined formulas. As of this writing, we support railway domain specific metrics including the average speed at which trains move in the railway network, the average amount of fuel consumed by the trains, amount of time trains spend waiting for signals to go green, and total distance traveled by the trains. These metrics are important to the railroad operators and serve as high-level system metrics for the experiment. Key operational metrics are used to help determine the health and efficiency of the railroad operation. By evaluating these metrics, researchers can assess the impact to the railroad system caused by each experiment or the cyber-attack simulation.

3.3.2 Domain Specific Customization. In addition to collecting individual simulation experiment metrics defined in the metrics library, it is often necessary to utilize a more sophisticated interpretation for gaining relevant insights. For example, a different unit of measurement may be necessary, or a more descriptive statistic may be developed through a combination of the smaller individual simulation metrics. Our domain specific customization module supports these efforts by providing the ability to define equations that relate the respective data collected by the metrics library.

4 CASE STUDY

For demonstration purposes, we utilize a scenario based on the Washington, D.C. Metro Railway System (see Fig. 3. This scenario builds a railway network mirroring to a part of the Washington, D.C. Metrol railway system coupled with signals (for stopping trains on the tracks or letting them proceed) and control switches (actuators for dynamically changing the tracks on which the trains will proceed). These railway switches are mirroring the functions of real-world railway junctions, where trains can be directed in different directions depending on their programmed destinations. The control of the actuator switches is determined by a control program that is executed based on the wireless signals that the approaching trains send to the control unit.

4.1 Attack Scenario

In this particular experiment, we focused on demonstrating the framework features of both the simulation testbed as well as the HIL testbed. For demonstrating cyber-attacks on the simulation, we utilize an integrity attack from the attack library. In particular, the integrity attack causes the railway switch to get incorrect messages from the approaching trains, thereby causing the trains to go on undesired tracks. Also, for demonstrating the cyber-attacks that are realizable realistically only in the physical hardware, we use the attack library's HIL-specific attacks. In particular, the DDOS attack is used to cause flood the network traffic destined to a specific railway signal controller. Consequently, the attacked railway signal controller is overwhelmed with the messages that it needs to process resulting in invalid or unchanged signal states. The result

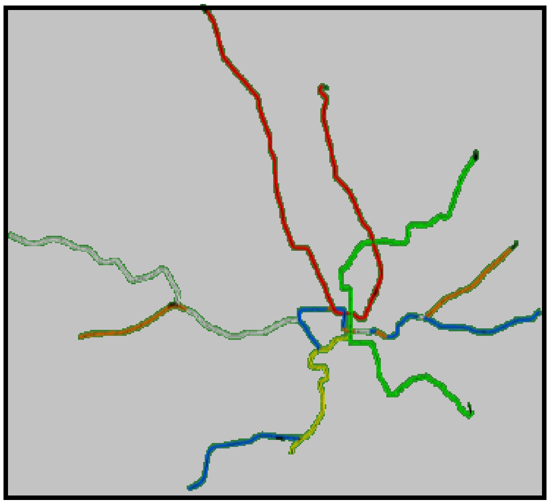


Figure 3: Washington, D.C. Metro Railway Network

of this is that the approaching trains get delayed until the railway signal updates to a green light. The experiment results from this scenario are described in the next subsection.

4.2 Results

To illustrate the results of our experimental scenario, we focus on analyzing the path of an individual train traveling through the Washington, DC metro network from Reston, VA to Greenbelt, MD. Utilizing the metrics library within our experiment design interface, we collect the following information: the average speed at which trains move in the railway network, the average amount of fuel consumed by the trains, amount of time trains spend waiting for signals to go green, and total distance traveled by the trains. Interestingly, the optimal path of the respective train is traveling east on the silver line, transferring to the blue line and traveling north of the inner city, and transferring to the green line to travel north east to Greenbelt. The attack process disrupts the train route through the following: the rail switch integrity attack results in the train being routed to the blue line south direction versus the optimal north route. As such, the train will have to travel around the southern perimeter of the city. Furthermore, while transferring to the green line, the DDOS attack on the corresponding rail signal causes a significant delay before the transfer is completed. Figures 4, 5, 6, and 7 illustrate the respective speed, fuel consumption, waiting time, and distance traveled for the train that experiences the worst impact due to attacks. In the figures, the red colored lines illustrate the train's metrics with the attack scenario enabled, while the blue colored line illustrates the train baseline scenario when traveling on the optimal path. The rail signal delay can be observed from approximately 9300 to 11000 seconds with the simulation. The figures show that there is a significant drop in speed at this time, the fuel consumption rate decreases as the train remains idle, and the accumulated waiting time increases sharply. Furthermore, the distance traveled metrics clearly show that the attack was successful as the train gets late in reaching its destination at Greenbelt by 1500 seconds.

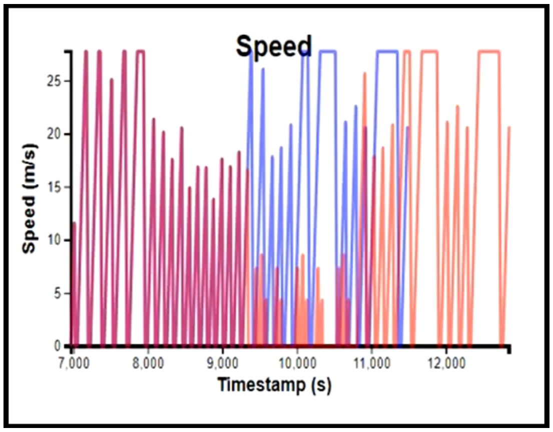


Figure 4: Speed

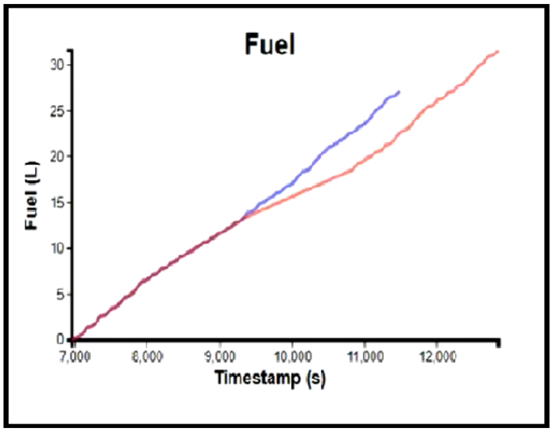


Figure 5: Fuel

5 RELATED WORK

With the significant increase in the utilization of digital technologies in the railway domain, including the European Railway Traffic Management System (ERTMS), is becoming more susceptible than ever to cyber-attacks [17]. The physical consequences of railway failures have been very publicly demonstrated in explosion-based attacks [24], as well as infrastructure failures [27]. Additionally, there has been much concern about the possibility of utilizing railway software for terrorist activities, serving as a prime vector to inflicting maximum damage to patrons who are not prepared for cyber-attacks [19]. As such, it is crucial to incorporate security and resilience requirements right from the design time, thereby maximizing trust in the safety of operations on their deployment into the field [14].

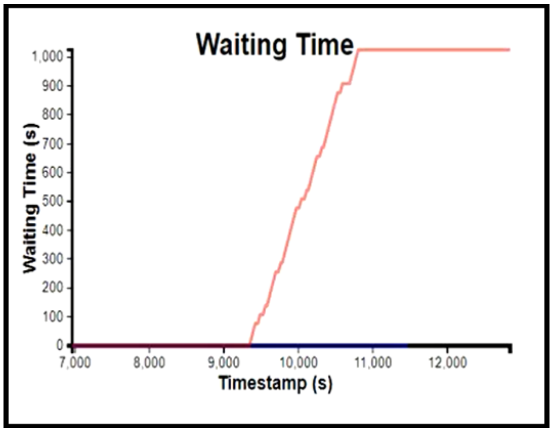


Figure 6: Waiting Time

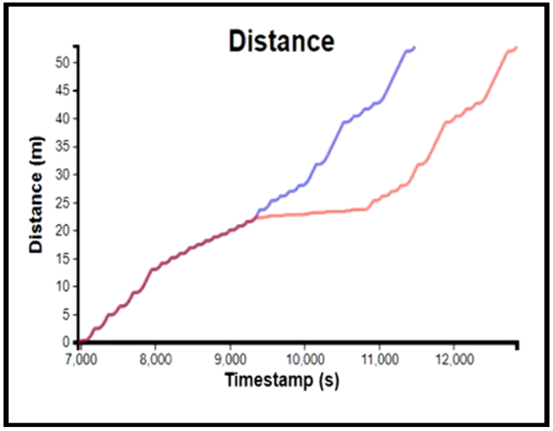


Figure 7: Distance

There has been increasing research in the railway industry focusing on the CPS perspective [20] including communication security, actuation safety [7], and train operator authentication [6]. Additionally, there has been a large amount of interest in the academic community in creating simulation testbeds for analyzing safety-critical CPS for security, safety, and resilience in the presence of cyber-attacks. The goal is to maintain predictable and safe operation during all scenarios, including when the system is under attack [5]. These frameworks have ranged from risk assessment tools [4], integrated simulation environments [8] [11], as well as transportation-specific attacker-defender modeling interfaces [15]. Additionally, the WebGME meta-modeling toolsuite has been a popular graphical user environment for easily and rapidly controlling CPS simulations [13].

6 CONCLUSION

In present day society, transportation is becoming more interconnected with the use of embedded devices for computation and control and wireless networking for communication. Railway is no exception, essentially becoming a computer on tracks. Even though there are significant benefits, including cost savings and an increase in travel efficiency, the introduction of electronic components and remote interfaces creates a significant avenue for attacker exploitation. With the safety-critical actuation of trains, locking down data from exfiltration is no longer enough as the railway operations also need to be maintained safely on a continual basis. A failure in this effort can lead to devastating consequences including high-speed train crashes, hazardous leaks of laboratory compounds or gasoline, and even loss of human life.

In this research work, we have illustrated a railway infrastructure experimentation platform, that can leverage cloud backend for large-scale computations, and that allows users to rapidly develop cyber-attack scenarios against train software to evaluate the resulting physical behavior. The goal of this approach is to allow railway designers to test the safety of algorithms and infrastructure well before deployment, providing trust in defense protections once in the field.

We developed three key components in our simulation framework, viz. the Experiment Controller, the Simulation Backend, and the Simulation Analyzer. The Simulation Backend provides the ability to evaluate railway infrastructure components in a simulation environment using the Veins simulator. Additionally, an integrated HIL testbed allows the users to emulate software on embedded microcontrollers similar to deployment environments. This allows for creating a realistic environment for getting a sense of the full spectrum of possible cyber-attack reactions, allowing for designers to have a more accurate sense of prioritizing the most critical vulnerabilities and attack vectors to protect against. Our Experiment Controller provides a graphical modeling environment for designing experiment scenarios including modeling and deployment of cyber-attacks. Further, we evaluated our platform with a case study of the Washington, DC metro network, where we captured domain-specific operational metrics as they were being generated, thereby validating our experimental design approach. We demonstrate the Simulation Analyzer that enables analysis of experiment results through integrated data capturing and visualization tools.

Further ahead, we plan to continue to extend our testbed for applying it to other transportation applications, including self-driving vehicles, and for enriching the re-configurable model libraries with more cyber-attacks and security solutions.

7 ACKNOWLEDGEMENTS

This work at Vanderbilt is supported by the National Security Agency (NSA) through award number H98230-18-D-0010, the National Institute of Standards and Technology (NIST) through award number 70NANB17H266, and the National Science Foundation (NSF) through award number CNS-1739328. No approval or endorsement of any commercial product by NSA, NIST, or NSF is intended or implied. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NSA, NIST, or NSF, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. This publication was co-authored by United States Government employees as part of their official duties and is, therefore, a work of the U.S. Government

and not subject to copyright. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSA, NIST, or NSF.

REFERENCES

[1] [n.d.]. Grafana - The open observability platform for metrics and analytics. https://grafana.com/. (Accessed on 01/21/2020).

[2] 2010. IEEE Std 1516–2010, IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)- Framework and Rules. , 3 pages.

[3] Andreas Bader, Oliver Kopp, and Michael Falkenthal. 2017. Survey and comparison of open source time series databases. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)-Workshopband* (2017).

[4] Bradley Potteiger, Goncalo Martins, and Koutsoukos, Xenofon. 2016. Software and attack centric integrated threat modeling for quantitative risk assessment. In *Proceedings of the Symposium and Bootcamp on the Science of Security*. ACM, 99–108.

[5] Bradley Potteiger, Zhenkai Zhang, and Xenofon Koutsoukos. 2018. Integrated instruction set randomization and control reconfiguration for securing cyber-physical systems. In *Proceedings of the 5th Annual Symposium and Bootcamp on Hot Topics in the Science of Security*. ACM, 5.

[6] Andrey V Chemov, Maria A Butakova, Ekaterina V Karpenko, and Oleg O Kartashov. 2016. Improving security incidents detection for networked multilevel intelligent control systems in railway transport. *Telfor Journal* 8, 1 (2016), 14–19.

[7] Jahanzeb Farooq and José Soler. 2017. Radio communication for communications-based train control (CBTC): A tutorial and survey. *IEEE Communications Surveys & Tutorials* 19, 3 (2017), 1377–1402.

[8] Himanshu Neema. 2018. Large-Scale Integration of Heterogeneous Simulations. *Ph.D. Dissertation, Vanderbilt University* (Jan. 2018).

[9] Himanshu Neema, Bradley Potteiger, Xenofon Koutsoukos, CheeYee Tang, and Keith Stouffer. 2018. Metrics-Driven Evaluation of Cybersecurity for Critical Railway Infrastructure. In *2018 Resilience Week (RWS)*. IEEE, 155–161.

[10] Himanshu Neema, Bradley Potteiger, Xenofon Koutsoukos, CheeYee Tang, and Keith Stouffer. 2018. Metrics-Driven Evaluation of Cybersecurity for Critical Railway Infrastructure. In *2018 Resilience Week (RWS)*. IEEE, 155–161.

[11] Himanshu Neema, Bradley Potteiger, Xenofon Koutsoukos, Gabor Karsai, Peter Volgyesi, and Janos Sztipanovits. 2018. Integrated Simulation Testbed for Security and Resilience of CPS. *The 33rd ACM/SIGAPP Symposium On Applied Computing* (Apr. 2018).

[12] Sidra Ijaz, Munam Ali Shah, Abid Khan, and Mansoor Ahmed. 2016. Smart cities: A survey on security concerns. *International Journal of Advanced Computer Science and Applications* 7, 2 (2016), 612–625.

[13] Gabor Karsai, Holger Krahn, Claas Pinkernell, Bernhard Rumpe, Martin Schindler, and Steven Völkel. 2014. Design guidelines for domain specific languages. *arXiv preprint arXiv:1409.2378* (2014).

[14] Sangjun Kim, Yuchang Won, In-Hee Park, Yongsoon Eun, and Kyung-Joon Park. 2019. Cyber-physical vulnerability analysis of communication-based train control. *IEEE Internet of Things Journal* 6, 4 (2019), 6353–6362.

[15] Xenofon Koutsoukos, Gabor Karsai, Aron Laszka, Himanshu Neema, Bradley Potteiger, Peter Volgyesi, Yevgeniy Vorobeychik, and Janos Sztipanovits. 2017. SURE: A modeling and simulation integration platform for evaluation of secure and resilient cyber–physical systems. *Proc. IEEE* 106, 1 (2017), 93–112.

[16] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. 2012. Recent Development and Applications of SUMO - Simulation of Urban MObility. *International Journal On Advances in Systems and Measurements* 5, 3&4 (December 2012), 128–138. http://elib.dlr.de/80483/

[17] Igor Lopez and Marina Aguado. 2015. Cyber security analysis of the European train control system. *IEEE Communications Magazine* 53, 10 (2015), 110–116.

[18] Neema, H., H. Nine, G. Hemingway, J. Sztipanovits, and G. Karsai. 2009. Rapid Synthesis of Multi-Model Simulations for Computational Experiments in C2. *Armed Forces Communications and Electronics Association - GMU Symposium, Critical Issues in C4I* (May 2009).

[19] M-Elisabeth Paté-Cornell, Marshall Kuypers, Matthew Smith, and Philip Keller. 2018. Cyber risk management for critical infrastructure: a risk analysis model and three case studies. *Risk Analysis* 38, 2 (2018), 226–241.

[20] Mouna Rekik, Christophe Gransart, and Marion Berbineau. 2018. Cyber-physical security risk assessment for train control and monitoring systems. In *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 1–9.

[21] Christoph Sommer, David Eckhoff, Alexander Brummer, Dominik S Buse, Florian Hagenauer, Stefan Joerer, and Michele Segata. 2019. Veins: The open source vehicular network simulation framework. In *Recent Advances in Network Simulation*. Springer, 215–252.

[22] Keith Stouffer, S Lightman, V Pillitteri, Marshall Abrams, and Adam Hahn. 2015. Guide to Industrial Control Systems (ICS) Security–NIST Special Publication (SP) 800-82 revision 2. *NIST, Tech. Rep* (2015).

[23] Andras Varga. 2019. A practical introduction to the OMNeT++ simulation framework. In *Recent Advances in Network Simulation*. Springer, 3–51.

[24] Wikipedia. 2020. 2004 Madrid train bombings — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=2004%20Madrid%20train%20bombings&oldid=934598985. [Online; accessed 21-January-2020].

[25] Wikipedia. 2020. 2015 Philadelphia train derailment — Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=2015%20Philadelphia%20train%20derailment&oldid=936015816. [Online; accessed 20-January-2020].

[26] Peng Zhang, Zsolt Lattmann, James Klingler, Sandeep Neema, and Ted Bapty. 2015. Visualization techniques in collaborative domain-specific modeling environment. In *SoutheastCon 2015*. IEEE, 1–6.

[27] Zhipeng Zhang, Xiang Liu, and Keith Holt. 2018. Positive Train Control (PTC) for railway safety in the United States: Policy developments and critical issues. *Utilities Policy* 51 (2018), 33–40.

Reference Datasets for Training and Evaluating RF Signal Detection and Classification Models

Timothy A. Hall, Raied Caromi, Michael Souryal, and Adam Wunderlich
Communications Technology Laboratory
National Institute of Standards and Technology
{tim.hall, raied.caromi, michael.souryal, adam.wunderlich}@nist.gov

***Abstract*—There are several potential uses of artificial intelligence (AI) and machine learning (ML) in next-generation shared spectrum systems. However, while recognized datasets exist in certain domains such as speech, handwriting and object recognition, there are no equivalent robust and comprehensive datasets in the wireless communications and radio frequency (RF) signals domain. Therefore, we propose creating an RF signal database focused on the signals, schemes, systems and environments found in these applications. Our intention is to create datasets that have the following characteristics: targeted around specific applications and systems relevant to the next generation of advanced communications, composed of traceable elements and carefully curated to be representative of the target application. Datasets may consist of field measurements, data collected in a controlled laboratory setting, synthetically generated waveforms or any combination of the above.**

***Index Terms*—RF dataset, deep learning, machine learning, detection, classification**

I. Introduction

Machine learning (ML) using deep learning (DL) algorithms has recently found considerable success in numerous research fields and applications as diverse as computer vision, speech recognition, image processing, language modeling and natural language processing, information retrieval, clustering, finance, robotics, business management, energy, transportation, and health care [1]–[5]. In line with the surge of interest in DL research, there are a number of potential uses of artificial intelligence (AI) in general, and ML in particular, in the next generation communications systems. For example, ML could be used to improve the performance of individual components of fifth generation (5G) cellular systems such as antenna configuration, beamforming training and tracking and overall multiple-input multiple-output (MIMO) system optimization. They could be used to model the environment via channel estimation and to control end-to-end performance by mapping quality of service (QoS) and quality of experience (QoE) measures to system parameters for given channel conditions. Prior to deployment, they could be used at the radio frequency (RF) planning stage to ensure sufficient coverage. In most of these communications applications, there is a need for the system to adapt to changing conditions, whether they be a wireless channel, end user traffic, or the service area of a provider [5]–[7]. Furthermore, ML and DL techniques have been used for cognitive radio systems and proposed for applications of physical layer and propagation models [8]–[10].

Another area of next generation communications that is becoming adaptive by nature is spectrum allocation—how much spectrum is used, by which users, and at what radio frequencies. To address increasing demand of a finite resource, spectrum allocation is moving away from fixed, exclusive use and towards dynamic sharing, where multiple, heterogeneous systems share the same spectrum without interfering with one another. Agencies such as the Defense Advanced Research Projects Agency (DARPA) are looking to AI and ML as enablers of efficient, automated spectrum sharing [11].

In shared spectrum systems, where two or more tiers or priority levels of users share a band, AI and ML can be used to increase spectrum utilization. One example with several applications for dynamic sharing of spectrum is the detection and classification of RF signals. Such capabilities enable identification of available spectrum and its use by secondary users (SUs) while protecting higher-priority incumbent users. More broadly, they can also aid in the detection of interfering signals for the purpose of enforcement.

Because research in RF signal classification requires data for training and testing classification algorithms, the availability of high quality datasets is vital. While recognized datasets exist in certain domains such as speech, handwriting and object recognition, researchers in academia and industry have observed that “no robust competition datasets exist in the emerging field of machine learning in the radio domain” [12].

Therefore, in order to facilitate and support the use of ML in next generation shared spectrum communications systems, we propose creating a curated RF signal database focused on the signals, schemes, systems and environments found in these applications. The result will be a publicly available database of RF signal measurements that covers a range of signal types (e.g., radar, 4G/5G cellular, GPS, telemetry), RF environments (or channel characteristics), signal-to-noise ratio (SNR) levels, types of interference and transceiver equipment.

In the following, we identify the technical and non-technical challenges of creating such a database in Section II. In Section III we outline the sources of data and the expected outcomes. After discussing related work in Section IV, we describe our approach to an RF signal database in Section V. We then describe an application-specific example of an RF signal database, that of incumbent radar waveforms in the shared 3.5 GHz band, and our initial efforts towards creating this database.

Hall, Timothy; Caromi, Raied; Souryal, Michael; Wunderlich, Adam. "Reference Data Sets for Training and Evaluating RF Signal Detection and Classification Models." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 14, 2019.

II. Challenges

Creating a curated RF signal database that can be used to train AI detection systems requires overcoming several technical challenges. Modern communication systems, such as long term evolution (LTE) systems, have thousands of configuration settings. In addition, many modern communication systems are adaptive, i.e., they change the underlying modulation and coding scheme, vary transmission power and schedule resources as needed in response to the changing RF channel and network loading. Creating datasets that are representative of these systems, as well as the full range of real-world conditions in which they operate, is critical to training generally useful ML models. This requires taking into account the effect of many variables. Identifying these variables, such as different types of transceiver hardware, configuration settings, network loading, and channel effects, is a significant aspect of the research. Perhaps more critical is determining which of these variables are the most important to training effective models. Although individual communications systems are carefully engineered, there is no body of knowledge or theory for constructing representative datasets from this high-dimensional space of non-stationary random processes.

Tagging or labeling the data is critical for training ML models. Examples include indicating whether the desired signal is active or not, SNR level, or identifying the signal or signals present. The most realistic environments come from field measurements, but these measurements are the most difficult to tag and label because ground-truth information is often not available.

There are also challenges that are not purely technical. For example, some waveforms, such as those from military systems, may be sensitive and restricted from public release. We need methods to obfuscate such waveforms (that is, remove sensitive aspects from them) in order to protect classified information while still facilitating the development of commercial detectors sharing the same spectrum.

III. Sources and Outcomes

Representative RF signal data can, in principle, be obtained from a variety of sources. These sources are described below, along with their particular advantages and disadvantages.

- *Field measurements*. Field measurements provide data samples from operational commercial or government equipment in real environments. Conducting measurement campaigns is time-consuming and offers limited to no control over system settings, network loading, and physical effects due to noise, interference and channel characteristics. For example, it is not possible to capture the signal from an operating radar over a controlled range of SNR values. Field measurements also present the tagging/labeling challenges detailed above.
- *Laboratory Testbed*. More control is possible using a laboratory testbed. In this case real transceiver equipment is used, but the measurements are collected in a laboratory environment with emulated network loading, and possibly, a channel emulator. In this setting, much greater control over configuration settings, network loading, signal levels, and channel effects is possible. Furthermore, accurate and complete tagging/labeling is easily accomplished. However, the representation of real-world RF environments may be limited.
- *Simulation*. In most domains, the use of simulated data for training ML models is discouraged. However, in some cases simulation may be useful to generate surrogate waveforms when the actual measured waveforms are unavailable.
- *Data from other organizations*. While much useful data can be collected and generated using the above methods, other organizations may have unique access to valuable data. For example, government agencies that use special-purpose RF equipment as part of their mission, e.g., the National Oceanic and Atmospheric Administration (NOAA), could be sources of signals that are otherwise difficult to obtain. Other organizations that perform spectrum monitoring for operational reasons or research may already possess a large amount of over-the-air recordings.

We envision these efforts will produce the following outputs that benefit the communications community and advance its application of ML and AI:

- A reference RF signal database that provides high quality, curated and traceable datasets for use in researching, developing, measuring and comparing ML models used to solve problems in next generation advanced communications systems.
- Research into how well surrogate waveforms, such as those for a classified radar, perform when used to train ML models. A surrogate waveform can be either an obfuscated version of a classified waveform, a similar non-classified radar or a synthetically generated waveform.
- Pioneering theory and practice in developing representative datasets of high-dimensional state space non-stationary random processes. Can techniques from experimental design used to reduce the testing state space be applied to the design of datasets? For example, the number of combinations of subcarrier modulation-coding scheme, e.g., in orthogonal frequency division multiplexing (OFDM), transmit power levels, channel effects, SNR, transceiver hardware and interference signal types and corresponding signal-to-interference ratio (SIR) is extremely large. Can training be done apart from an exhaustive enumeration of all combinations of these factors? Another question is whether we can mix synthetic signals with those collected from field or lab measurements in the same dataset in order to cover a needed range and combination of variables.

IV. Related Work

In his essay "50 Years of Data Science" [13], Donoho identifies the Common Task Framework (CTF) as the key to the success of predictive modeling. The first of the three elements of a successful CTF instance he cites is a "publicly available training dataset involving, for each observation, a list of (possibly many) feature measurements, and a class label for that observation." As we noted in the introduction, recognized public datasets exist in many domains. One example is the modified National Institute of Standards and Technology (MNIST) handwritten digit database [14].

While recognized datasets like those mentioned above have greatly advanced the usefulness of ML models in their respective domains, there are no equivalents for the field of RF signals. Some researchers and programs have made partial datasets available, but these are a byproduct of their research and not complete or mature enough to be of general use. We discuss this research below.

In [12], O'Shea et al. construct datasets for classifying the modulation type of a received radio signal. They argue that while, in general, the use of synthetic or simulated data for the learning phase of ML is not recommended, radio communication signals are an exception. Most of the steps in a communications system are synthetic and deterministic, i.e., modulation and coding. Those that are not, such as channel effects, can be well characterized and modeled mathematically. The authors constructed datasets using several modulation schemes, including binary phase shift keying (BPSK), quadrature phase shift keying (QPSK), 8-phase shift keying (8PSK) and 16-quadrature amplitude modulation (16QAM). They used a matched filter to obtain a baseline performance result. The datasets are available for download [15]. Data are stored in Hierarchical Data Format 5 (HDF5) format [16] as complex floating point values representing baseband in-phase/quadrature (IQ) values with 2 million examples, each 1024 samples long.

While O'Shea et al. argue for the appropriateness of using synthetic data for developing ML models for communications systems, de Vrieze et al. show that real-world signals and conditions are important [17]. They chose two particular real-world examples: heterogeneous transceiver hardware (i.e., low-grade universal software radio peripherals (USRPs) and high-grade lab instruments), and in-band interference from a modulated signal. The authors describe the database of signal samples they created. Each element of the dataset is 1 ms of captured data sampled at 50 MHz. From each element a feature vector of 12 features, statistical and spectral, was computed. The modulation schemes used were BPSK, QPSK, 16QAM, 64QAM for both the desired signal and co-channel interference and OFDM for the desired signal only. The signals were transmitted over a short distance in a laboratory environment. Thus, realistic channel effects were not considered. The datasets are available online, though they contain only the features that the authors used in the experiments in the paper, not IQ samples.

V. Approach

Our intention is to create datasets that have the following characteristics:

- Targeted around specific applications and systems relevant to the next generation of advanced communications
- Composed of traceable elements that follow the Findable, Accessible, Interoperable and Reusable (FAIR) guiding principles for scientific data management and stewardship [18]
- Carefully curated to be representative of the target application

We discuss each of these characteristics in more detail below. First, the datasets are targeted in that we do not consider aspects of a communications system or application in isolation or combined in arbitrary ways, and we do not aim to be exhaustive, covering all existing schemes. For example, the previous work cited above focused on identifying a limited set of modulation and coding schemes not tied to a particular standard or system. While such datasets may be useful for initial, proof-of-concept investigations in ML for detection and classification, they are of limited use in solving practical research problems. Datasets that are directly relevant to a specific RF band, system, or standard with unresolved technical challenges are more likely to be useful.

Second, datasets should be traceable. All elements of the RF dataset must be identified and described in the documentation. This is in addition to any tagging or labeling. Specifically, the following must be documented for any dataset.

- A globally unique and persistent identifier, FAIR [18].
- Date, time and location of collection.
- Bandwidth and sample rate at time of collection.
- Type of data collected. This refers to the categories of field measurements, laboratory measurements or simulated (synthetic) signals.
- Signal(s). These could be radar signals, communication signals or some other type. The signal can be collected from an over-the-air transmission, conducted measurements, or synthetically generated. If it is synthetically generated, then the code used to generate it should be identified and made publicly available. Everything needed to replicate the signal must be available. There may be multiple signals in a single dataset, and depending on the application, a given signal may be considered to be the desired signal or an interfering signal.
- Transceiver equipment. This is the actual transmitter or receiver hardware used to send and receive the desired signal. It can be, for example, commercial hardware, lab-grade instruments or software-defined radios (SDRs). de Vrieze et al. demonstrated the importance of including real hardware effects from heterogeneous hardware in the training set [17]. For both the transmitter and receiver, the configuration should be specified, e.g., with configuration files, along with any software, including version number. In general, for a collected waveform, the receiver is the measurement instrument, so any measurement uncertainty should be specified, as well.
- Channel. The channel is not only the medium of transmission but more broadly the effects upon the signal during transmission. Thus, a channel could be an actual, physical path through which the transmission propagates

or it could be a set of channel propagation effects that operate on the signal via a channel emulator.

- Noise. In general these are components of the received signal other than the desired signal or an interfering signal. Noise usually follows some sort of probability distribution, such as additive white Gaussian noise (AWGN), and is present at a defined SNR.
- Processing of collected data. Any processing applied to the collected data must be included. This includes operations such as band filtering, down-conversion, downsampling and normalizing.
- Lineage. If this dataset is a subset of another one or has been derived from it in any other manner, e.g., downsampled, both the parent dataset's universally unique identifier and any operations applied to it must be identified.
- Any additional metadata that may be available, e.g., from diagnostic monitoring software.

The third quality we require is curation of the datasets. This means the careful selection of communications systems and applications and their documentation. It also includes a uniform format for storage and organization of the data and documentation files and careful management to keep a portion of the dataset sequestered for validation of models. More broadly, however, it addresses the more difficult problem of selecting a range of signals, equipment, channel, SNR levels and interference types necessary to have a representative dataset. As mentioned in Section I, there is no existing body of work on how to do this for RF signals, or for other problems of such high-dimensional state space. Developing theory and practice of how to do this is an important open research problem.

VI. RF Dataset Example: 3.5 GHz Radar Waveforms

Building on our extensive experience with measurements, standardization, and testing of systems in the shared 3.5 GHz Citizens Broadband Radio Service (CBRS) band in the U.S. [19], an initial dataset of interest to us is for the detection of incumbent radar signals in this band. In this application, a Navy shipborne radar is the incumbent operator. Secondary users are LTE operators, who can use the band when the radar is not present and active in their region of operation. Because the incumbent does not directly signal to the SUs when it is active, the SUs must rely on a network of detectors in order to determine the presence of radar and thus when to stop transmitting.

There are several advantages to starting with an application that we have experience with, namely radar detection in the CBRS band. The first is that we understand the problem well and have identified some of the relevant issues. We know the system performance requirements and have also been actively involved in associated measurement campaigns. Extensive experience with an application aids the curator in making the dataset representative and appropriately challenging for, in this case, detection and classification. For example, in the CBRS band, there are two primary sources of interference to a radar detector. The first is emissions from commercial CBRS systems that will be co-channel with the radar. The second is the out-of-band emissions (OOBE) of adjacent-band radars spilling into the CBRS band. Our prior work in support vector machine classifiers of the current radar in the band revealed the importance of including both types of interference in the evaluation [20]. As illustrated in the receiver operating characteristic (ROC) curves of Fig. 1, a peak analysis classifier (PAC) is more robust to the OOBE than a higher-order statistic classifier (HSC). While the HSC performs comparably to the PAC in the presence of co-channel commercial LTE emissions, it breaks down when subjected to OOBE. Not including samples with OOBE in the dataset would give a misleading indication of relative classifier performance.

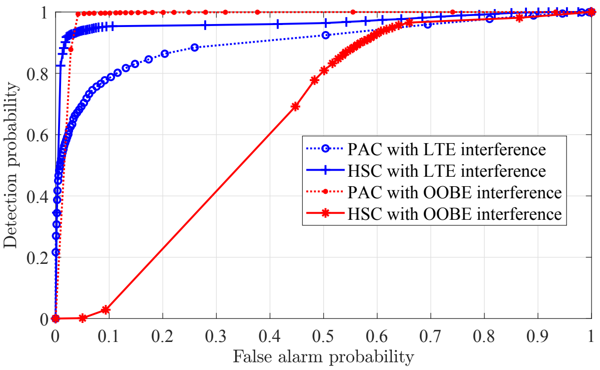


Figure 1. Two classifiers with two types of interference. Data taken from Fig. 5 and Fig. 6 in [20]

The current incumbent radar operating in the 3.5 GHz CBRS band is an unmodulated pulse radar of which we have made extensive field measurements [21], [22]. However, at the time of writing, these measurements are classified and not available for public release. Furthermore, more sophisticated radars are expected to be deployed in the band for which we currently have no field or laboratory measurements. As a result, a dataset of 3.5 GHz radar waveforms will have to rely on synthetically generated (simulated) waveforms, at least initially. Fortunately, because we have access to field measurements of the current radar, we can easily test the effectiveness of simulated waveforms against actual waveforms. For example, one could attempt to demonstrate that a classifier trained with simulated waveforms performs well with field-measured waveforms. Alternatively, it may be possible to match certain characteristics of simulated waveforms, e.g, power distributions of the signal, noise, and interference, to measured data.

We have begun to develop a dataset of simulated 3.5 GHz radar waveforms that will be made available to the public. The waveforms are derived from five categories of pulse radar that are representative of current and future systems in the 3.5 GHz band [23]. They consist of unmodulated and frequency/angle modulated radar signals with a range of values for operating parameters such as pulse width and pulse repetition rate. Each parameter has a defined range of values with a specific step size that ensures a minimum change in value. The waveforms are generated by randomly selecting the values of these parameters from the discrete ranges given in [23, Table 1]. In addition, other parameters for the waveforms—such as SNR, power level of the radar signal and noise or interference, sampling frequency, and waveform duration—are carefully chosen to render a representative and useful dataset. Finally, each waveform is saved with the full set of parameters that were used during the generation process. The inclusion of these parameters is important for both analyzing the performance of a classifier against a specific parameter as well as providing the traceability of each waveform.

The success of commercial deployment in the CBRS band in the U.S. may motivate the use of similar spectrum sharing techniques in additional bands. For example, the National Telecommunications and Information Administration (NTIA) is currently studying the 3450 MHz to 3550 MHz band (3.4 GHz band) for potential commercial wireless broadband use [24]. In the U.S., military radar systems also operate in this band. Therefore, a dataset developed for the 3.5 GHz CBRS band is potentially applicable to the 3.4 GHz band, as well.

VII. Summary

In summary, we propose creating an RF signal database focused on the signals, schemes, systems and environments found in next generation spectrum sharing systems. Datasets will be targeted around specific applications and systems, composed of traceable elements and carefully curated to be representative of the target application. We plan to use field measurements, data collected in a controlled laboratory setting, synthetically generated waveforms or any combination of the above.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[2] L. Deng and D. Yu, *Deep Learning: Methods and Applications*. Hanover, MA, USA: Now Publishers Inc., 2014.

[3] Y. Li, “Deep reinforcement learning,” *CoRR*, vol. abs/1810.06339, 2018. [Online]. Available: https://arxiv.org/abs/1810.06339

[4] L. Liu, W. Ouyang, X. Wang, P. W. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, “Deep learning for generic object detection: A survey,” *CoRR*, vol. abs/1809.02165, 2018. [Online]. Available: https://arxiv.org/abs/1809.02165

[5] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. Chen, and L. Hanzo, “Thirty years of machine learning: The road to pareto-optimal next-generation wireless networks,” *CoRR*, vol. abs/1902.01946, 2019. [Online]. Available: https://arxiv.org/abs/1902.01946

[6] C. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” *IEEE Communications Surveys Tutorials*, pp. 1–1, 2019.

[7] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, “Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks,” *CoRR*, vol. abs/1710.02913, 2017. [Online]. Available: https://arxiv.org/abs/1710.02913

[8] M. Bkassiny, Y. Li, and S. K. Jayaweera, “A survey on machine-learning techniques in cognitive radios,” *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1136–1159, 2013.

[9] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 4, pp. 563–575, Dec 2017.

[10] M. Ribero, R. W. Heath Jr., H. Vikalo, D. Chizhik, and R. A. Valenzuela, “Deep learning propagation models over irregular terrain,” in *Proc. IEEE ICASSP*, May 2019.

[11] P. Tilghman, “AI will rule the airwaves: A DARPA grand challenge seeks autonomous radios to manage the wireless spectrum,” *IEEE Spectrum*, vol. 56, no. 6, pp. 28–33, June 2019.

[12] T. J. O’Shea and J. Corgan, “Convolutional radio modulation recognition networks,” *CoRR*, vol. abs/1602.04105, 2016. [Online]. Available: http://arxiv.org/abs/1602.04105

[13] D. Donoho, “50 years of data science,” *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 745–766, 2017. [Online]. Available: https://doi.org/10.1080/10618600.2017.1384734

[14] The MNIST Database of handwritten digits. [Online]. Available: http://yann.lecun.com/exdb/mnist

[15] Rf datasets for machine learning. [Online]. Available: https://www.deepsig.io/datasets

[16] The HDF Group. [Online]. Available: https://www.hdfgroup.org

[17] C. de Vrieze, L. Simic, and P. Mahonen, “The importance of being earnest: Performance of modulation classification for real RF signals,” in *2018 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Oct 2018, pp. 1–5.

[18] FAIR principles. [Online]. Available: https://www.go-fair.org/fair-principles

[19] “Citizens broadband radio service,” 47 C.F.R. § 96, 2016.

[20] R. Caromi and M. Souryal, “Detection of incumbent radar in the 3.5 GHz CBRS band using support vector machines,” in *Proc. IEEE Sensor Signal Processing for Defence (SSPD) Conference*, May 2019.

[21] P. Hale, J. Jargon, P. Jeavons, M. Lofquist, M. Souryal, and A. Wunderlich, “3.5 GHz radar waveform capture at Point Loma,” *NIST TN 1954*, May 2017. [Online]. Available: https://doi.org/10.6028/NIST.TN.1954

[22] ——, “3.5 GHz radar waveform capture at Fort Story,” *NIST TN 1967*, October 2017. [Online]. Available: https://doi.org/10.6028/NIST.TN.1967

[23] F. H. Sanders, J. E. Carroll, G. A. Sanders, R. L. Sole, J. S. Devereux, and E. F. Drocella, “Procedures for laboratory testing of environmental sensing capability sensor devices,” National Telecommunications and Information Administration, Technical Memorandum TM 18-527, Nov. 2017. [Online]. Available: http://www.its.bldrdoc.gov/publications/3184.aspx

[24] “NTIA identifies 3450-3550 MHz for study as potential band for wireless broadband use,” by David J. Redl, Assistant Secretary for Communications and Information and NTIA Administrator, Feb. 2018. [Online]. Available: https://www.ntia.doc.gov/blog/2018/ntia-identifies-3450-3550-mhz-study-potential-band-wireless-broadband-use

Hall, Timothy; Caromi, Raied; Souryal, Michael; Wunderlich, Adam. "Reference Data Sets for Training and Evaluating RF Signal Detection and Classification Models." Paper presented at 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, US. December 09, 2019 - December 14, 2019.

Developing Models for a 0.8 mm Coaxial VNA Calibration Kit within the NIST Microwave Uncertainty Framework

Jeffrey A. Jargon[1], Christian J. Long[1], Ari Feldman[1], and Jon Martens[2]

[1]National Institute of Standards and Technology, 325 Broadway, M/S 672.03, Boulder, CO 80305 USA
[2]Anritsu Company, Morgan Hill, CA 95037 USA
Email: jeffrey.jargon@nist.gov, Tel: +1.303.497.4961

Abstract — **We developed models for a 0.8 mm coaxial vector network analyzer (VNA) calibration kit within the NIST Microwave Uncertainty Framework. First, we created physical models of commercially-available standards and included error mechanisms in each of the standards' constituent parameters that were utilized to propagate uncertainties. Next, we calibrated a network analyzer with this calibration kit and compared measurements and uncertainties of two verification devices with data provided by the manufacturer. We found the measurements agreed to within their respective uncertainties.**

Index Terms — **calibration, coaxial, physical models, uncertainty, vector network analyzer, verification.**

I. Introduction

With the demand for faster data-transfer rates, the wireless industry continues to develop technology in the millimeter-wave (30-300 GHz) spectrum. One potential impediment is the connector interface. Typically, waveguides are used at higher frequencies since they are easier to manufacture and less lossy than coaxial lines. The main disadvantage of waveguides, however, are their limited bandwidths, precluding broadband frequency coverage. As an alternative, 0.8 mm coaxial connectors have recently been developed that provide uninterrupted coverage up to 145 GHz [1-2].

Researchers at the National Institute of Standards and Technology (NIST) are actively pursuing research in millimeter-wave frequencies, including small- and large-signal network analysis, modulated-signal characterization, over-the-air (OTA) testing for advanced cellular applications, antenna metrology, channel measurements and modeling [3], and mismatch correction for electro-optic sampling (EOS).

The EOS system at NIST is the United States' primary standard for high-speed waveform calibration of photodiodes and is traceable to the SI through fundamental physics. A photodiode calibrated using this system serves as a time- and frequency-domain transfer standard and allows for subsequent calibrations of high-speed oscilloscopes, light-wave component analyzers, comb generators, and high-speed modulated signals [4]. To calculate the electrical waveform at the photodiode's coaxial connector from the voltage measured in the coplanar waveguide (CPW) by the EOS system, a change in reference plane is required. This requires vector network analysis to characterize the reflection coefficients of the photodiode and on-wafer resistor, as well as the scattering-parameters (*S*-parameters) of the probe head. Currently, the EOS system is limited to 110 GHz due to the 1.0 mm coaxial connectors on the photodiode and probe head. If these two devices were fitted with 0.8 mm coaxial connectors and characterized with correlated uncertainties, the EOS system could provide calibrated photodiodes at frequencies up to 145 GHz.

With extending the EOS capabilities as one of our prime motivators, we utilized the NIST Microwave Uncertainty Framework (MUF) [5] to develop physical models of a commercially-available 0.8 mm coaxial calibrations kit. The MUF utilizes parallel sensitivity and Monte-Carlo analyses, and enables us to capture and propagate the significant *S*-parameter measurement uncertainties and statistical correlations among them [6]. By identifying and modeling the physical error mechanisms in the calibration standards, we can determine the uncertainties in *S*-parameters, including their cross-frequency correlations. These uncertainties can then be propagated to measurements of devices under test. In the following sections, we describe our methodology in further detail, and compare measurements and uncertainty estimates made on two verification devices with data provided by the manufacturer.

II. Model Development

Our commercial 0.8 mm calibration kit consisted of two sets of standards – male and female offset opens, offset shorts, and loads (OSL) for frequency coverage up to 80 GHz, and three pairs of offset shorts (SSS) with differing lengths for frequency coverage between 80 GHz and 145 GHz [7]. The manufacturer included dimensions and uncertainties for the offset lengths and inner and outer conductor diameters, as well as polynomial models for each of the standards.

Using the MUF, we constructed our own physical models of the calibration standards with closed-form expressions for coaxial lines of finite metal conductivity [8]. The MUF was also used for automatically propagating the uncertainties to the calibrated verification devices in conjunction with the calibration engine, StatistiCAL [9-10], which utilizes a "mix-and-match" philosophy to VNA calibrations.

In our previous work with Type-N, 3.5 mm, and 2.4 mm coaxial calibrations standards, we have compared our physical models to measurements of the standards performed with an independent multiline thru-reflect-line (TRL) calibration [11-12]. Since we do not have access to any 0.8 mm airlines, we were unable to do so here. However, we did compare our physical models to the manufacturer's polynomial models and

Jargon, Jeffrey; Long, Chris; Feldman, Ari; Martens, Jon. "Developing Models for a 0.8 mm Coaxial VNA Calibration Kit within the NIST Microwave Uncertainty Framework." Paper presented at 94th ARFTG Microwave Measurement Conference, San Antonio, TX, US. January 26, 2020 - January 29, 2020.

optimized our inductances and capacitances to match the frequency responses of the manufacturer's standards, especially with regards to the phase delays.

The low-band (< 80 GHz) OSL standards were modeled with the values and uncertainties (standard errors) listed in Tables I and II. Table I lists the physical error mechanisms related to the inner and outer conductors and pins of the offset transmission lines, and Table II lists the physical error mechanisms specific to the pairs of opens, shorts, and loads. The high-band (80 GHz to 145 GHz) offset-short (SSS) standards were modeled with the values and uncertainties listed in Tables I and III. Table III lists the physical error mechanisms specific to the offset shorts. The thru connection was modeled as a zero-length transmission line and included error mechanisms.

Table I. Physical error mechanisms of the 0.8 mm calibration standards.

Mechanism (units)	**Value ± Uncertainty**
Inner Conductor Diameter (mm)	0.347 ± 0.003
Outer Conductor Diameter (mm)	0.8 ± 0.003
Pin Diameter (mm)	0.2 ± 0.003
Pin Depth (mm)	-0.010 ± 0.005
Metal Conductivity (S/m)	$6\times10^6 \pm 5\times10^6$
Relative Dielectric Constant	1.000535 ± 0
Dielectric Loss Tangent	0 ± 0

Table II. Physical error mechanisms of the 0.8 mm OSL standards.

Mechanism (units)	**Value ± Uncertainty**
OPEN	
Male Offset Length (mm)	1.2 ± 0.013
Male Capacitance (pF)	0.012 ± 0.002
Male Conductance (1/Ω)	0 ± 0
Female Offset Length (mm)	1.2 ± 0.013
Female Capacitance (pF)	0.0124 ± 0.002
Female Conductance (1/Ω)	0 ± 0
SHORT	
Male Offset Length (mm)	1.2 ± 0.013
Male Inductance (nH)	0.0042 ± 0.002
Male Resistance (Ω)	0 ± 0.1
Female Offset Length (mm)	1.2 ± 0.013
Female Inductance (nH)	0.0042 ± 0.002
Female Resistance (Ω)	0 ± 0.1
LOAD	
Male Resistance (Ω)	50.2 ± 0.1
Male Inductance (nH)	0.006 ± 0.002
Male Shunt Capacitance (pF)	0.0001 ± 0
Female Resistance (Ω)	50.2 ± 0.1
Female Inductance (nH)	0.010 ± 0.002
Female Capacitance (pF)	0.0001 ± 0

Table III. Physical error mechanisms of the 0.8 mm SSS standards.

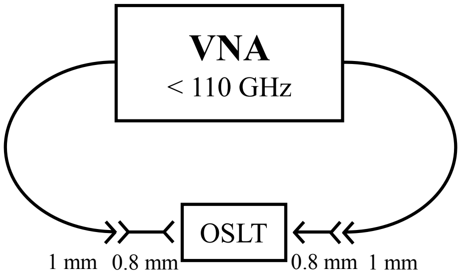
Mechanism (units)	**Value ± Uncertainty**
SHORT 1	
Male Offset Length (mm)	1.2 ± 0.013
Male Inductance (nH)	0.005 ± 0.002
Male Resistance (Ω)	0 ± 0.1
Female Offset Length (mm)	1.2 ± 0.013
Female Inductance (nH)	0.0057 ± 0.002
Female Resistance (Ω)	0 ± 0.1
SHORT 2	
Male Offset Length (mm)	1.63 ± 0.013
Male Inductance (nH)	0.007 ± 0.002
Male Resistance (Ω)	0 ± 0.1
Female Offset Length (mm)	1.63 ± 0.013
Female Inductance (nH)	0.00535 ± 0.002
Female Resistance (Ω)	0 ± 0.1
SHORT 3	
Male Offset Length (mm)	2.06 ± 0.013
Male Inductance (nH)	0.002 ± 0.002
Male Resistance (Ω)	0 ± 0.1
Female Offset Length (mm)	2.06 ± 0.013
Female Inductance (nH)	0.0005 ± 0.0005
Female Resistance (Ω)	0 ± 0.1

III. Measurement Comparison

Figure 1 illustrates our VNA measurement configuration. One VNA, equipped with 1 mm test ports, enabled us to make measurements at frequencies up to 110 GHz (200 MHz spacing, 20 Hz IF bandwidth, and -10 dBm output power), and the other VNA, equipped with WR-8 test ports, allowed us to measure at frequencies between 90 GHz and 140 GHz (100 MHz spacing, 10 Hz IF bandwidth, and -10 dBm output power). Both VNAs were equipped with 0.8 mm adapters (Port 1 being female and Port 2 being male). Even though our OSL standards (low-band) were specified by the manufacturer up to only 80 GHz, we used them to 110 GHz since our VNA equipped with 1 mm test ports had the capability to measure these frequencies. Our SSS standards (high-band) were measured from 90 GHz to 140 GHz using the VNA equipped with WR-8 test ports. Thus, the two calibrations overlapped between 90 GHz and 110 GHz.

Figures 2-5 show calibrated *S*-parameters and corresponding confidence bounds calculated for the two verification devices – a Beatty line and a matched airline. The dark red curves correspond to the manufacturer's values using the OSLT calibration up to 80 GHz, and the light red curves correspond to the manufacturer's values using the SSST calibration from 80 GHz to 140 GHz. The dark blue curves correspond to our measured values using the OSLT calibration up to 110 GHz, and the light blue curves correspond to our measured values using the SSST calibration from 90 GHz to 140 GHz. The dashed curves correspond to confidence bounds, in our case at 95% intervals.

In Figures 2-5, we see that our measurements appear to be noisier than those provided by the manufacturer. This may be in part due to our frequency spacing being 100 or 200 MHz and the manufacturer's being 1 GHz. Our confidence intervals were also smaller in general but did not include a repeatability component. The manufacturer stated their confidence intervals were "extremely conservative" and "include a large budget for cable flex and connector repeatability."



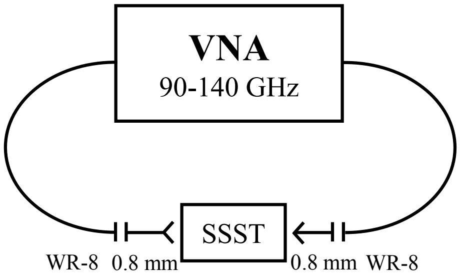


Figure 1. VNA measurement configuration.

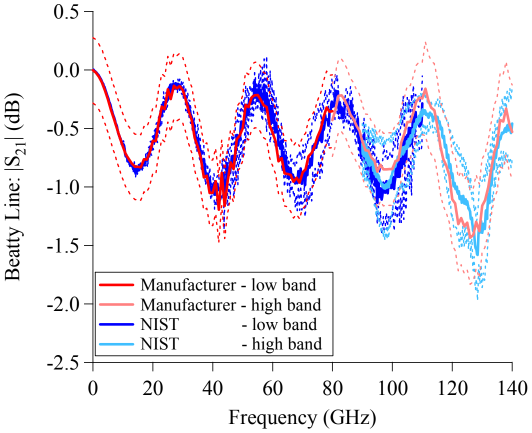


Fig. 2. Measurements and confidence intervals of the Beatty line's transmission coefficients.

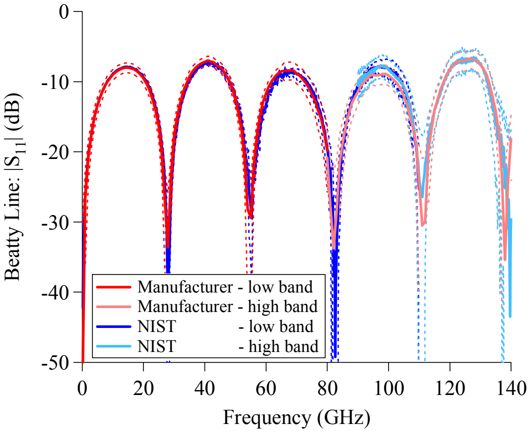


Fig. 3. Measurements and confidence intervals of the Beatty's lines reflection coefficients.

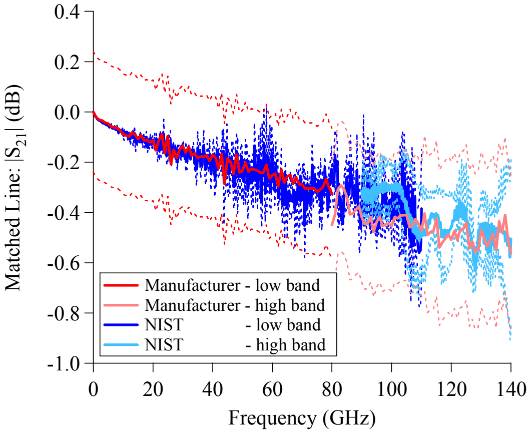


Fig. 4. Measurements and confidence intervals of the matched line's transmission coefficients.

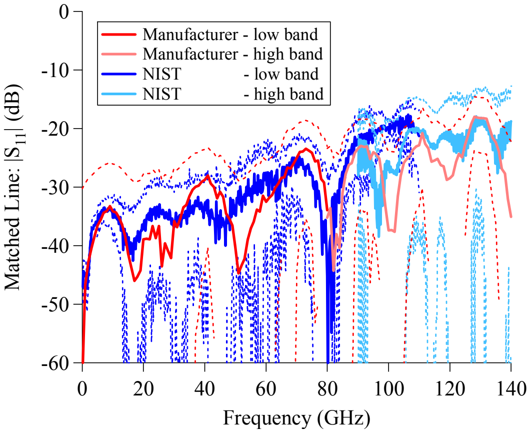


Fig. 5. Measurements and confidence intervals of the matched line reflection coefficients.

Nonetheless, the variations between our measurements and the manufacturer's agreed to within their respective confidence intervals at most frequencies. Some discrepancies may be attributed to connector repeatability, and the different calibration algorithms and standard definitions. Additionally, our measurements on the two VNAs agreed to within their respective confidence intervals for the overlapping frequencies between 90 GHz and 110 GHz, as illustrated in Figures 6-7, which zoom in on the areas from Figures 2 and 4, respectively.

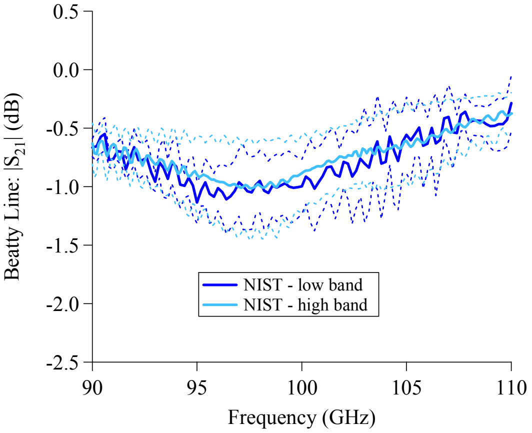


Fig 6. Comparing NIST measurements of the Beatty line's transmission coefficients from 90 GHz to 110 GHz.

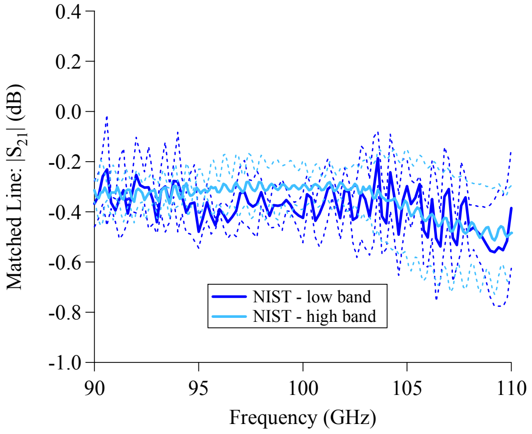


Figure 7. Comparing NIST measurements of the matched line's transmission coefficients from 90 GHz to 110 GHz.

IV. Conclusions

We have developed physical models of a 0.8 mm coaxial calibration kit for vector network analyzers that support calibration algorithms within the NIST Microwave Uncertainty Framework. After calibrating a network analyzer with this calibration kit and comparing measurements and uncertainties of two verification devices with data provided by the manufacturer, we found the measurements agreed to within their respective uncertainties.

Future work in this area includes a study of connector repeatability and developing an approach for achieving traceability. Traceability would require that the OSL standards be measured with a more fundamental calibration, such as multiline TRL or SSST, where dimensional tolerances are traceable to NIST. Then, the resulting measurements from the fundamental calibration would be used as models for the OSL standards.

References

[1] C. Tumbaga, "0.8 mm Connectors Enable D-Band Coaxial Measurements," *Microwave Journal*, pp. S6-S12, Mar. 2019.

[2] T. Roberts, Y. S. Lee, J. Martens, and W. Oldfield, "Development and Traceable Measurements of 0.800 mm Coaxial Airline: Broadband Standard to 145 GHz," *81st ARFTG Microwave Measurement Conference*, Jun. 2013.

[3] K. A. Remley, J. A. Gordon, D. Novotny, A. E. Curtin, C. L. Holloway, M. T. Simons, R. D. Horansky, M. S. Allman, D. Senic, M. Becker, J. A. Jargon, P. D. Hale, D. F. Williams, A. Feldman, J. Cheron, R. Chamberlin, C. Gentile, J. Senic, R. Sun, P. B. Papazian, J. Quimby, M. Mujumdar, and N. Golmie, "Measurement Challenges for 5G and Beyond," *IEEE Microwave Magazine*, pp. 41-56, Jul./Aug. 2017.

[4] P. D. Hale, D. F. Williams, and A. Dienstfrey, "Waveform Metrology: Signal Measurements in a Modulated World," *Metrologia*, vol. 55, pp. S135-S151, 2018.

[5] D. F. Williams, NIST Microwave Uncertainty Framework, www.nist.gov/services-resources/software/wafer-calibration-software, 2019.

[6] A. Lewandowski, D. F. Williams, P. D. Hale, C. M. Wang, and A. Dienstfrey, "Covariance-Matrix-Based Vector-Network-Analyzer Uncertainty Analysis for Time- and Frequency-Domain Measurements," *IEEE Trans. Microwave Theory Tech.*, vol. 58, no. 7, pp. 1877-1886, July 2010.

[7] Anritsu, "3659 0.8 mm Calibration/Verification Kit and 2300-580-R System Performance Verification Software," User Guide, Sep. 2016.

[8] A. Lewandowski, "Multi-Frequency Approach to Vector-Network Analyzer Scattering-Parameter Measurements," Ph.D. Thesis, Warsaw University of Technology, 2010.

[9] D. F. Williams, StatistiCAL VNA Calibration Software Package, www.nist.gov/services-resources/software/wafer-calibration-software, 2019.

[10] D.F. Williams, C.M. Wang, and U. Arz, "An Optimal Vector-Network-Analyzer Calibration Algorithm," *IEEE Trans. Microwave Theory and Tech.*, vol. 51, no. 12, pp. 2391-2401, Dec. 2003.

[11] J. A. Jargon, C. H. Cho, D. F. Williams, and P. D. Hale, "Physical Models for 2.4 mm and 3.5 mm Coaxial VNA Calibration Kits Developed within the NIST Microwave Uncertainty Framework," *85th ARFTG Microwave Measurement Conference*, May 2015.

[12] J. A. Jargon, D. F. Williams, and P. D. Hale, "Developing Models for Type-N Coaxial VNA Calibration Kits within the NIST Microwave Uncertainty Framework," *87th ARFTG Microwave Measurement Conference*, May 2016.

UE-to-Network Relay Discovery in ProSe-enabled LTE Networks

Samantha Gamboa, Alexandre Moreaux, David Griffith, and Richard Rouil
National Institute of Standards and Technology
Gaithersburg, MD, USA
{samantha.gamboa, alexandre.moreaux, david.griffith, richard.rouil}@nist.gov

***Abstract*—The UE-to-Network Relay functionality was introduced to Long Term Evolution (LTE) cellular networks by the 3rd Generation Partnership Project (3GPP) in Release 13. In this technology, User Equipment (UEs) acting as Relay UEs are used to extend network coverage to cell-edge and out-of-coverage Remote UEs. One important part of this functionality is direct discovery, which is used by the Remote UEs willing to reach the network to detect the Relay UEs in proximity that can provide the desired connectivity service. In this paper, we study this protocol considering both discovery models defined in the LTE standard, and we develop analytical models to characterize the average time a Remote UE takes to discover a Relay UE using each discovery model. We validate the analytical models using system level simulations and we study the sensitivity of the metrics to different parameters of the protocol and number of UEs involved in the UE-to-Network Relay discovery.**

***Index Terms*—LTE, D2D, ProSe, UE-to-Network Relay, Discovery, Network Modeling**

I. INTRODUCTION

The 3rd Generation Partnership Project's (3GPP) Long Term Evolution (LTE) Proximity Services (ProSe) technology allows User Equipment (UEs) to communicate on a Device-to-Device (D2D) basis, directly sending information to one another (if the range permits it) via a direct link known as the sidelink (SL). This technology not only allows direct communication in areas within the network coverage, where an eNodeB (eNB) could coordinate SL resource allocation, but it also enables out-of-coverage UEs to communicate using the SL, in which case autonomous resource allocation is used based on pre-configured parameters [1].

One important feature of the ProSe technology is the UE-to-Network Relay functionality, where in-coverage UEs (Relay UEs) can act as network relays, redirecting traffic to and from another UE (Remote UE) in proximity of the network. Using UE-to-Network Relay UEs to extend coverage is critical to Public Safety users, especially in emergency scenarios where communication between intervening team members and incident command stations should not be interrupted. Thus, team members equipped with ProSe-enabled UEs can act as Relay UEs and ensure service continuity to out-of-coverage or cell-edge team members whose UEs will act as Remote UEs.

In order to reach the network, a Remote UE should search for Relay UEs in proximity using a ProSe direct discovery procedure, select the most suitable one, and connect to it using the one-to-one ProSe direct communication procedure [2]. In this paper, we focus on the ProSe direct discovery procedure for UE-to-Network Relay.

Both ProSe direct discovery models defined in the 3GPP standard (A and B) can be used for UE-to-Network Relay discovery [3]. In Model A, Relay UEs periodically broadcast *announcement* messages to advertise their presence and the connectivity service they can provide. Remote UEs actively listen for those messages. In Model B, the procedure is initiated by the Remote UE, which broadcasts *solicitation* messages with the connectivity service it is looking for. Listening Relay UEs that provide the solicited service will then send a *response* message. The received discovery announcements (Model A) or responses (Model B), and their associated signal strengths, are then used by the Remote UE to conduct the Relay UE selection and start the one-to-one ProSe direct communication procedure.

ProSe-enabled UEs use the Physical Sidelink Discovery Channel (PSDCH) to broadcast discovery messages on the SL. The UEs are provisioned with a pool of discovery resources to use, which repeats periodically in time. The resource allocation within each pool can be either *network-assisted*, i.e., the eNodeB persistently schedules the resources to be used by the UEs, or *UE-selected*, where each UE selects randomly the resources to be used.

In this paper, we focus on the UE-selected allocation, as eNodeB scheduling information may not be always available for the Remote UEs, e.g., when they are out-of-coverage. The UE-selected allocation brings a risk of collision interference if multiple UEs pick the same resources for a given transmission, which may result in Relay UEs not being discovered by a Remote UE even though they are in proximity. To alleviate this issue, a transmission probability was defined in the 3GPP standard as part of the discovery resource pool, which is used by each transmitter UE to decide whether to transmit a discovery message in a given discovery period. We developed an analytical framework to quantify the time taken by a Remote UE of interest to discover any Relay UE and also a given Relay UE in proximity depending on the discovery pool parameters and the discovery model used by the UEs. We validated the theoretical models using system level simulations performed in our ns3 ProSe module described in [4] and enhanced with the UE-to-Network Relay functionality.

The rest of the paper is organized as follows. We discuss

Gamboa Quintiliani, Samantha; Moreaux, Alexandre; Griffith, David W.; Rouil, Richard A. "UE-to-Network Relay Discovery in ProSe-enabled LTE Networks." Paper presented at International Conference on Computing, Networking and Communications (ICNC 2020), Big Island, HI, US. February 17, 2020 - February 20, 2020.

prior works related to ProSe direct discovery and the contributions of this paper in Section II. In Section III, we describe the models that allow us to obtain the time a Remote UE takes to discover any Relay UE and a given Relay UE in proximity. In Section IV, we provide numerical results to validate the models and we discuss the protocol sensitivity to the model parameters. Finally, Section V summarizes our contributions and discuss future work.

II. Related Work

As discussed in the previous section, the direct discovery resource allocation can be either network-assisted or UE-selected. Early works have focused on network-assisted direct discovery. In [5], Xenakis et al. provide an analytical model to determine the probability that two UEs in the network detect each other using a D2D link and provide a sensitivity analysis considering several network parameters such as transmission power and eNodeB density. In [6] and [7], the authors rely on centralized scheduling schemes to avoid collisions and expedite the discovery process. These studies assume full knowledge of the network and deviate from the 3GPP standard procedures in place for ProSe direct discovery.

The following works address direct discovery with UE-selected allocation. In [8], the authors present a model based on stochastic geometry and use it to calculate how many UEs can be discovered in a given number of discovery periods considering channel conditions. Bagheri et al. consider similar metric in [9], and propose that UEs randomly select their transmit power to alleviate interference in the discovery channel. Both models consider that the UEs are sending discovery messages every discovery period, disregarding the transmission probability mechanism.

Li and Liu proposed an alternative to the transmission probability to avoid collision interference [10]. Instead of the UEs deciding each period if they should transmit based on the transmission probability, they randomly decide in which discovery period to transmit within a set of successive periods.

In [11], Griffith and Lyons developed an analytical model used to obtain the optimal transmission probability for a given discovery resource pool configuration and number of UEs performing discovery. This optimal transmission probability minimizes the time required for a successful discovery message transmission, and near-optimal performance can be achieved when rounding it up to the next higher multiple of 1/4 to be consistent with the values in the 3GPP standard. These results are used in [12] to develop an adaptive algorithm in which the UEs adjust their transmission probabilities depending on the number of discovered UEs over time.

In this paper, we build upon the model in [11] and generalize it to differentiate between Remote UEs and Relay UEs. We also complete it to determine the average time a Remote UE needs to find an unspecified Relay UE. Additionally, we characterize Model B discovery using state machines as has been explored by Griffith et al. in [13], obtain equivalent performance metrics, and compare both models performances for a given discovery pool configuration. Finally, we validate the models using system level simulations of the actual standard protocol implementation.

III. Analytical Model

The notation we use in the paper is summarized in Table I.

A. System Model

1) Scenario: We consider a group of N_x Remote UEs (G) and a group of N_y Relays UEs (H) deployed without prior knowledge of the area. All the devices are in each other's respective ranges, X is a given Remote UE and Y a given Relay UE; both randomly chosen. We consider that devices are either Remote or Relay UEs (i.e., $G \cap H = \varnothing$).

2) The Discovery Resource Pool: Resource allocation will happen by UEs choosing their own resources in the PSDCH discovery resource pool independently from each other. The PSDCH resource pool is a periodical grid in the time-frequency plane composed of Physical Resource Blocks (PRB) that we model as a $N_f \cdot N_t$ matrix as depicted in Fig. 1. Each row corresponds to a PRB pair and each column to a subframe set. Each resource is a single transport block composed of a pair of adjacent PRB that occupy the same subframe. Any UE wishing to transmit is to generate a uniformly distributed random value $p_1 \in [0, 1]$ that it compares with a given threshold value denoted by *txProbability*. The 3GPP standard defines that *txProbability* can take the values 0.25, 0.50, 0.75, or 1.00. The discovery message is sent if $p_1 <$ *txProbability*. If successful, the UE picks a resource in the pool with uniform probability. Let $\theta = P(p_1 \leqslant \textit{txProbability})$ for the rest of this paper.

3) Metrics of interest: This paper focuses on two metrics: The average time a given Remote UE of interest needs to find:

- A given Relay UE, and
- Any Relay UE, referred to as First Relay UE.

We will establish analytical models for these two metrics for Model A and B. We elected a probabilistic approach and looked into the behaviors of our system in a given PSDCH period and deduced how many periods are necessary on average to complete the discovery for each case. For the rest of the paper, let $Z_1 \rightarrow Z_2$ be the event 'UE Z_1 successfully sends UE Z_2 a message', $Z_1 o Z_2$ be 'UE Z_1 discovers UE Z_2', $Z?$ be 'UE Z successfully sends a message to a Relay/Remote UE' and $Z \sim$ be 'UE Z discovers a Relay/Remote UE'. Note that Z, Z_1, and Z_2 can designate interchangeably a Relay or Remote UE, depending on the model in context.

4) Final Considerations: The following assumptions were made for the development of the analytical models presented in the next sections:

- We assume that all UEs belong to the same security domain and are authorized to perform UE-to-Network Relay discovery. This approach could be deemed Open Access although its long-term applications are focused around Restricted Access in the Public Safety context.
- We assume a worst-case transmission scenario were the devices other than our devices of interest are permanently trying to send messages.

Gamboa Quintiliani, Samantha; Moreaux, Alexandre; Griffith, David W.; Rouil, Richard A. "UE-to-Network Relay Discovery in ProSe-enabled LTE Networks." Paper presented at International Conference on Computing, Networking and Communications (ICNC 2020), Big Island, HI, US. February 17, 2020 - February 20, 2020.

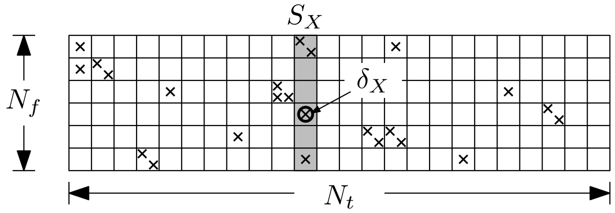


Fig. 1. The discovery resource pool model, showing the transmissions of various UEs and indicating the location of X's discovery message δ_X and the set of subframes it occupies S_X [11, Fig. 3].

TABLE I
LIST OF SYMBOLS

Symbol	Definition
$P(A)$	Probability of event A
G	The set of Remote UEs in context
N_x	card(G)
UE X	Randomly chosen Remote UE of interest from G
H	The set of Relay UEs in context
N_y	card(H)
UE Y	Randomly chosen Relay UE of interest from H
δ_X	Discovery message sent by UE X
S_X	Set of subframes occupied by δ_X
N_r	Number of resources in discovery pool
N_f	Number of PRB pairs in discovery pool
N_t	Number subframe sets in discovery pool
θ	Probability that a given UE transmits
T	Markov state transition matrix
P_1,P_2	Illustration probability functions
N	T's Fundamental Matrix
Z, Z_1, Z_2	Arbitrary UEs ($\in G \cup H$)
$Z_1 \to Z_2$	$'Z_1$ successfully sends a message to Z_2'
$Z_1 o Z_2$	$'Z_1$ discovers Z_2'
$Z?$	$'Z$ successfully sends a message to a Relay/Remote UE$'$
$Z \sim$	$'Z$ discovers a Relay/Remote UE$'$

- We assume that senders other than our UEs of interest do not stop sending discovery messages after receiving responses from other devices. This condition guarantees us a stable environment to work in.
- The half-duplex effect prevents devices from transmitting and listening simultaneously on the SL. Although it is a factor that could impact the discovery when using Model B, we assume the Remote UEs transmit solicitations every other discovery period. This allows that the following period is dedicated to the reception of the Relay UE responses.
- Lastly, this paper will neglect processing times completely. As soon as there is an intent to transmit, the involved device attempts to do so. These delays are probably not as significant as control channel loss probability and other phenomena but do still exist.

B. Relay Discovery Model A

1) Given Relay UE: In Model A, given the absence of transmission from Remote UEs and the systematic one of all Relay UEs, XoY is equivalent to $X \to Y$ and furthermore to 'Y's discovery message does not collide'. By Bayes' Theorem on the universe $\{\{\text{`}Y \text{ emits'}\}, \{\text{`}Y \text{ does not emit'}\}\}$:

$$P(X \to Y) = P(\text{`}Y \text{ emits'})P(\text{`}Y\text{'s message does not collide'}) \tag{1}$$

with $P(\text{`Y emits}') = \theta$.

To determine P('Y's message does not collide') , we condition on how many of the other $N_y - 1$ Relay UEs transmit, which has a binomial distribution with probability mass function $f(k; N_y - 1; \theta) = \binom{N_y-1}{k}\theta^k(1-\theta)^{N_y-1-k}$. The probability that a Relay UE other than Y picks the same resource if it transmits successfully is $\frac{1}{N_r}$. By applying the Binomial Theorem, we get:

$$P(XoY) = \theta\left(1 - \frac{\theta}{N_r}\right)^{N_y-1}. \tag{2}$$

<u>$N.B$</u>: This method is reminiscent of the one used in [11].

2) First Relay UE: A discovery ($X \sim$) is equivalent to a successful transmission ($X?$) in Model A. The only way $X?$ does not happen is if all transmissions collide. By conditioning on the number of devices that transmit and using Bayes' Theorem:

$$P(X?) = \sum P(\text{`at least one request does not collide'} \\ |\text{`k requests are emitted'})P(\text{`k requests are emitted'}). \tag{3}$$

Let C_k be the event 'All of k requests collide'. Determining $P(C_k)$ now comes down to a classic balls in bins problem where we look for the number of combinations leaving no ball alone in a box:

$$P(C_k) = \sum_{i=0}^{min(k,N_r)} (-1)^i \binom{N_r}{i}\binom{k}{i} i! \left(\frac{1}{N_r}\right)^i \left(1 - \frac{i}{N_r}\right)^{k-i}. \tag{4}$$

$Proof$: The probability that no ball is alone in a box is $P(C_k) = 1 - P(\text{`at least one ball alone'})$. Let ν be the number of balls that are alone after k balls are randomly distributed into N_r bins. We can define a set of k events, $S = \{A_i\}_{i=1}^{k}$ such that event A_i occurs if the ith ball is alone in a box. Jordan's formula gives the probability that at least j events in the set S occur, which is

$$P(\nu \geqslant j) = \sum_{i=j}^{j} (-1)^{i-j}\binom{i-1}{j-1} B_i, \tag{5}$$

where $B_i = \sum_{1 \leqslant m_1 \leqslant m_2 \leqslant \ldots \leqslant m_i \leqslant k} P(A_{m_i} \cap \ldots \cap A_{m_i})$ is the ith binomial moment of ν, which here corresponds to the probability of balls numbered $m_1, ..., m_i$ being alone in i bins. We compute this probability by taking the ratio of the number of ways that k distinguishable balls can be arranged so that the balls with the index values $m_1, m_2, \ldots, m_i$ are alone in i distinguishable bins to the number of ways that k balls can be arranged in N_r bins. The former is the product of the number of ways to choose i bins out of N_r bins, the number of ways to arrange the i isolated balls among the i chosen bins and the number of ways to arrange the remaining $(k - i)$ balls among the remaining $(N_r - i)$ bins (note that this can include cases where some of these balls end up alone, but we are concerned only with balls $m_1, m_2, \ldots, m_i$) whereas the latter is simply N_r^k. We now have

$$\begin{aligned} P(A_{m_i} \cap ... \cap A_{m_i}) &= \binom{N_r}{i} i! \frac{(N_r - i)^{k-i}}{N_r^k} \\ &= \binom{N_r}{i} \frac{i!}{N_r^i} \left(1 - \frac{i}{N_r}\right)^{k-i}. \end{aligned} \tag{6}$$

We can now deduce

$$B_i = \binom{k}{i} \binom{N_r}{i} \frac{i!}{N_r^i} \left(1 - \frac{i}{N_r}\right)^{k-i}. \tag{7}$$

And finally

$$\begin{aligned} &P(C_k) = 1 - P(\nu \geqslant 1) \\ &= \sum_{i=0}^{min(k,N_r)} (-1)^i \binom{N_r}{i} \binom{k}{i} i! \left(\frac{1}{N_r}\right)^i \left(1 - \frac{i}{N_r}\right)^{k-i}. \end{aligned} \tag{8}$$

This results leads us to:

$$\begin{aligned} &P(X \sim) = \sum_{k=1}^{N_y} \binom{N_y}{k} \theta^k (1-\theta)^{N_y - k} \\ &\sum_{i=0}^{min(k,N_r)} (-1)^{i+1} \binom{N_r}{i} \binom{k}{i} i! \left(\frac{1}{N_r}\right)^i \left(1 - \frac{i}{N_r}\right)^{k-i}. \end{aligned} \tag{9}$$

3) Deducing desired metrics: For both these models, given that the resource selection process is independent in distinct periods, the number of PSDCH periods to accomplish the aforementioned events has a geometric distribution. We can conclude that the mean time (in number of periods) for each of the scenarios is the inverse of the probabilities, given by:

$$t(XoY) = \frac{1}{P(XoY)} \tag{10}$$

and

$$t(X \sim) = \frac{1}{P(X \sim)}. \tag{11}$$

N.B: In model A, the probability of X discovering a given group of Relay UEs is equal to the one where all the Remote UEs discover the same Relay UEs, given that messages are broadcasted to all Remote UEs in range.

C. Relay Discovery Model B

With Model B, other Remote UEs than X will be using and competing for resources. Thus, N_x will now logically appear in the models and all N_x Remote UEs we take into account are active. The approach we used for Model A cannot lead us to the time metric for Model B as the geometric distribution criteria is no longer fulfilled. We elected Markov models as the most appropriate to handle Model B. In fact, the sojourn times of the states in our model do not exactly follow the geometric distribution, which is a requirement for Markov chains. Nevertheless, using a semi-Markov chain adds considerable complexity to the model for an incremental gain in accuracy and does not bring any significant insight to this paper. Therefore, we can reasonably approximate the sojourn times and continue using standard Markov chains.

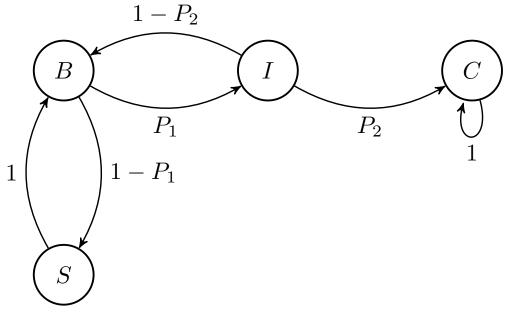


Fig. 2. Markov chain modeling the system when using discovery Model B.

We propose to study a simplified version of Model B, were PSDCH pools are divided into solicitation and response periods, meaning that Remotes UEs would attempt to send solicitations only every other period and that Relay UEs would have the following period to try responding. The discovery is now achieved when a successful solicitation and response happen in any pair of neighboring PSDCH periods. Any of the Remote UEs in G can trigger a response that allows X to discover the concerned Relay UE. This simpler approach still depicts Model B's behaviors and metrics appropriately, whilst neglecting timer expirations and the half-duplex effect.

1) The Model: The Markov chain depicted in Fig. 2 models our system when using discovery Model B and is used for obtaining both metrics of interest. We define P_1 to be the probability that a Remote UE's solicitation is successfully received by a Relay UE, and define P_2 to be the probability that a Relay UE's response is successfully received by a Remote UE in the general case. In our model, the Remote UE starts in the Beginning state, B, and progresses to the Intermediary state, I, if a Relay UE receives its solicitation. Otherwise the Remote UE moves to sleeping state S (as the current period is not a Remote UE transmission one) and then returns to state B to attempt transmission once again. If the Relay UE's response reaches the Remote UE, then it progresses to the Completion state, C, which is the sole absorbing state in the Markov chain.

The chain in Fig. 2 has the following transition matrix:

$$T = \begin{bmatrix} 0 & P_1 & 1 - P_1 & 0 \\ 1 - P_2 & 0 & 0 & P_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{12}$$

That we can partition likewise:

$$T = \begin{bmatrix} Q & R \\ \mathbf{0} & 1 \end{bmatrix}. \tag{13}$$

This chain's fundamental matrix defined by $N = (1-Q)^{-1}$ with $Q = T_{[1;3][1;3]}$ can be determined directly by the cofactor method:

$$N = \frac{1}{P_1 P_2}\begin{bmatrix} 1 & P_1 & 1-P_1 \\ 1-P_2 & P_1 & (1-P_1)(1-P_2) \\ 1 & P_1 & 1-P_1(1-P_2) \end{bmatrix}. \quad (14)$$

By definition, the average time to reach the absorbing state (in periods) when starting at state B is:

$$t(B \to C) = \sum_{j=1}^{3} N_{1,j} = \frac{2}{P_1 P_2}. \quad (15)$$

For both our metrics P_1, i.e., the probability of a successful solicitation, is essentially the same as Model A's first Relay UE model with the Remote UEs being the transmitters. This is the case because any of the Remote UEs can solicit a specific (or any) Relay UE for it to answer X, our Remote UE of interest. Given our assumptions, we can be certain of the solicitation reaching our Relay UE of interest if it does not collide, and therefore

$$P_1 = P(X?) = \sum_{k=1}^{N_x} \binom{N_x}{k} \theta^k (1-\theta)^{N_x-k} \sum_{i=1}^{min(k,N_x)} (-1)^{i+1} \binom{N_r}{i} \binom{k}{i} i! \left(\frac{1}{N_r}\right)^i \left(1-\frac{i}{N_r}\right)^{k-i}. \quad (16)$$

The response however does vary depending on the desired metric. The given Relay UE Model will have one Relay UE competing with others for resources to reach X, whereas the first Relay UE Model requires any Relay UE to reach X.

a) Given Relay UE: For the Given Relay UE model, we simply have:

$$P_2 = P(Y \to X) = \theta \left(1 - \frac{\theta}{N_r}\right)^{N_y - 1}, \quad (17)$$

which leads to:

$$t(XoY) = \frac{2}{P(Y \to X)P(X?)}. \quad (18)$$

b) First Relay UE: For the First Relay UE model, the response can be issued by any Relay UE, leading to:

$$P_2 = P(Y?) = \sum_{k=1}^{N_y} \binom{N_y}{k} \theta^k (1-\theta)^{N_y-k} \sum_{i=1}^{min(k,N_y)} (-1)^{i+1} \binom{N_r}{i} \binom{k}{i} i! \left(\frac{1}{N_r}\right)^i \left(1-\frac{i}{N_r}\right)^{k-i}, \quad (19)$$

and finally:

$$t(X \sim) = \frac{2}{P(Y?)P(X?)}. \quad (20)$$

IV. Numerical Results

A. System Level Simulations

We validated the models using system-level simulations. We used the ns-3 ProSe model described in [4] enhanced to support the UE-to-Network Relay functionality. We deployed N_y Relay UEs and N_x Remote UEs in proximity and all UEs are configured with the same discovery pool parameters (N_f, N_t, *txProbability*). The discovery period was set to 320 ms which is the minimum value defined in the standard. All UEs start the direct discovery at the same time. Relay UEs send discovery announcement messages on every period when using Model A, and Remote UEs send discovery solicitations every other period when using Model B. We considered ideal channel conditions and discovery messages sent by multiple UEs in the same resources are dropped to be consistent with the analytical models. Message recovery depending on Signal to Interference plus Noise Ratio (SINR) and error models will be addressed in future work. The metrics of interest were calculated for a given Remote UE, which was chosen randomly in each trial. In the simulations, we consider that a Remote UE discovers a Relay UE when it successfully receives an announcement (Model A) or response (Model B) message from that Relay UE. We performed 1000 independent trials for each configuration, and all results are presented using the mean values and 95 % confidence intervals for each metric.

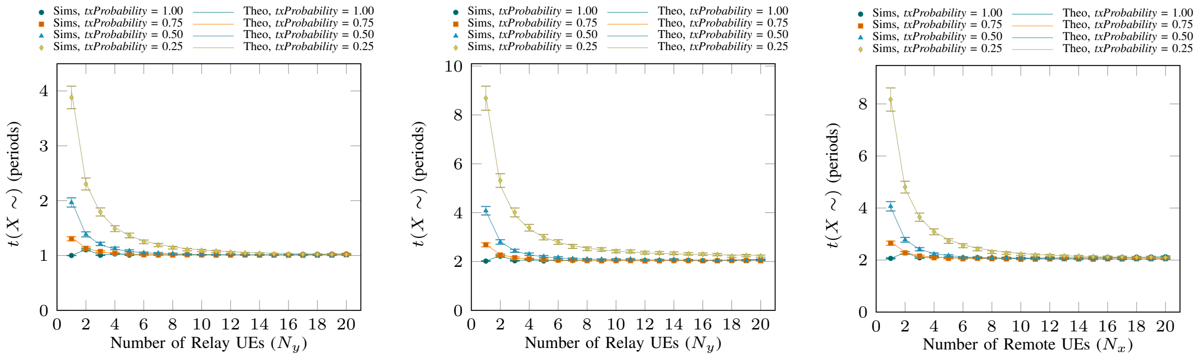
B. Results Discussion

The results shown in Fig. 3 and Fig. 4 were generated using a resource pool of $N_f = 2$ PRBs and $N_t = 5$ subframes (SFs), and we present results for the four values of *txProbability* defined in the 3GPP standard $(0.25, 0.50, 0.75, 1.00)$. We observe a close agreement between the theoretical and the system level simulations results, validating the accuracy of our models.

The results illustrate the effect of the number of UEs contending for resources in the discovery time for the Remote UE of interest. On the one hand, we observe in Fig. 3(a) and Fig. 3(b) that $t(X \sim)$ decreases when N_y increases, as the probability of at least one Relay UE choosing a non-colliding discovery resource increases with N_y. The scenario with $N_y = 1$ and *txProbability* = 1.00 is of course the exception, as there is no contention for resources and the discovery messages are received in every discovery period. On the other hand, Fig. 4(a) and Fig. 4(b) show that $t(XoY)$ increases with N_y, as the number of UEs contending for discovery resources increases, the probability that the Relay UE of interest choose a non-colliding discovery resource decreases, thus delaying the discovery.

While Fig. 3(b) and Fig. 4(b) shows results for $N_x = 10$ Remote UEs, we observed similar trends depending on N_y for different values of N_x when using Model B.We do not show these results due to space limitation. Please note that in this evaluation, the results for Model A are independent of N_x, as Remote UEs are only listening for announcements and all Remote UEs receive them at the same time.

Fig. 3(c) and Fig. 4(c) show the trend when varying N_x for Model B and $N_y = 20$ Relay UEs. The discovery time

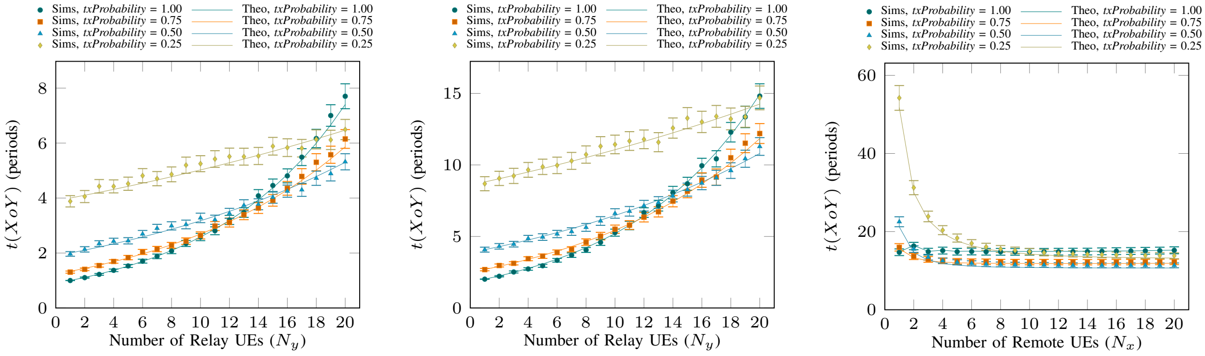


(a) Model A.

(b) Model B, N_x = 10 Remote UEs.

(c) Model B, N_y = 20 Relay UEs.

Fig. 3. Average number of discovery periods needed by a Remote UE to discover the first Relay UE ($t(X \sim)$). Mean and 95 % confidence intervals are shown for the system level simulation results.



(a) Model A.

(b) Model B, N_x = 10 Remote UEs.

(c) Model B, N_y = 20 Relay UEs.

Fig. 4. Average number of discovery periods needed by a Remote UE to discover a given Relay UE ($t(XoY)$). Mean and 95 % confidence intervals are shown for the system level simulation results.

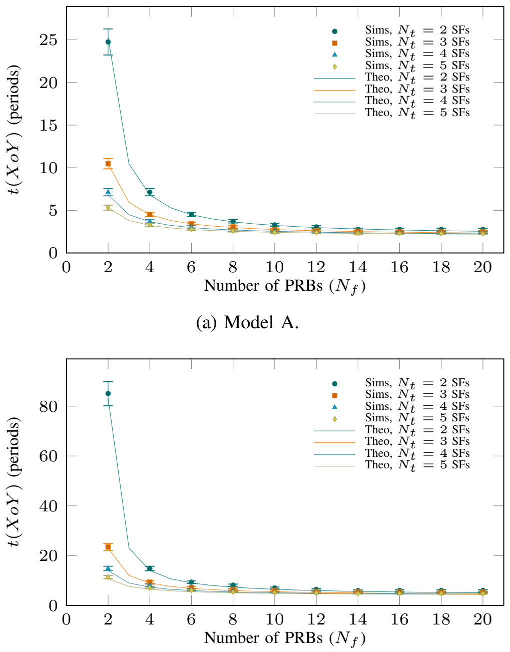
decreases when N_x increases because the probability that at least one solicitation is sent in a non-colliding discovery resource increases with the number of Remote UEs sending solicitations (N_x). Then, the discovery time is affected by all N_y Relay UEs trying to respond to this solicitation in the following discovery period, similar to Model A behavior with $N_y = 20$. Similar trends were observed for other values of N_y and results are not shown due to space limitations.

Fig. 3 and Fig. 4 also allow us to see the effect of *txProbability* on the discovery time. When the number of UEs contending for resources is small, the largest discovery times are observed in all plots for *txProbability* = 0.25. When N_y and N_x increases, the benefits of reducing concurrent transmissions with a lower transmission probability become more evident; as seen in Fig. 4 the value of *txProbability* that minimizes the discovery time varies depending on N_y and N_x. These observations are consistent with similar evaluations made for group member discovery in [11].

We observe that the discovery time with Model B is longer than with Model A when using equivalent pool configuration. This is expected, as the Relay UE discovery depends on the successful reception of two messages by different UEs when using Model B, i.e., the solicitation from the Remote UE and the response from the Relay UE. Thus, the average discovery time is at least doubled when using Model B. However, when looking at different pool configurations and for low number of UEs, Model B can provide lower discovery times than Model A. For example, if we consider $N_y = 4$ Relay UEs, Fig. 4(a) shows that for Model A the average discovery time for *txProbability* = 0.25 is 4.32 discovery periods, while in Fig. 4(b) the given Relay UE can be discovered in 2.77 discovery periods for *txProbability* = 1.00, or in 3.41 discovery periods for *txProbability* = 0.75 when using Model B. Although we have focused our analysis on time metrics, it is important to keep in mind that the difference in the dynamics of the discovery models impacts other metrics such as energy consumption, which may balance out the extra time needed by Remote UEs when using Model B.

In Fig. 5 we illustrate the impact of the resource pool size on the discovery time for both models. Increasing N_f

Gamboa Quintiliani, Samantha; Moreaux, Alexandre; Griffith, David W.; Rouil, Richard A. "UE-to-Network Relay Discovery in ProSe-enabled LTE Networks." Paper presented at International Conference on Computing, Networking and Communications (ICNC 2020), Big Island, HI, US. February 17, 2020 - February 20, 2020.



(a) Model A.

(b) Model B.

Fig. 5. Average number of discovery periods needed by a Remote UE to discover a given Relay UE ($t(XoY)$). Parameters: *txProbability* = 0.50, $N_y = 20$ Relay UEs and $N_x = 20$ Remote UEs. Mean and 95 % confidence intervals are shown for the system level simulation results.

impacts favorably the discovery time, but we observe that the benefits become negligible for larger values of N_f. Increasing N_t provides considerable reductions in the discovery time. However, increasing the discovery resource pool size may come to the expense of reducing the time and frequency resources available for other ProSe operations.

Our analytical models and system level simulator could be used by network operators to select a suitable configuration depending on the targeted performance, expected deployment, and use cases. A typical use case example in the public safety context is a group of First Responders working on an incident in an area with partial network coverage. In this scenario, N_y units stay in the in-coverage area and their devices act as Relay UEs. The other N_x units move to the out-of-coverage area and their devices act as Remote UEs. In this context, operators can use our models to search for the combination of N_f, N_t, and *txProbability* parameters that minimizes the discovery time, i.e., $t(XoY)$ or/and $t(X\sim)$, for the N_x out-of-coverage units, and thus minimize the impact on their service continuity. As the number of units responding to an incident is highly variable, operators can use the models to find a configuration that can provide acceptable results for the largest number of expected N_x and N_y combinations. Furthermore, dynamic algorithms that choose suitable configurations depending on local conditions could help to further improve the discovery performance and are currently open research topics.

V. Conclusions and Future Work

The UE-to-Network Relay discovery time in UE-Selected mode is affected by the discovery pool configuration, the amount of UEs contending for discovery resources, and the discovery model used by the ProSe-enabled UEs. In this paper, we developed a model to quantify the average time taken by a Remote UE of interest to discover any Relay UE and a given Relay UE in proximity depending on those parameters. We validated the model using system level simulations and we have shown the sensitivity of the protocol to the discovery model, number of UEs participating on the discovery, and pool configurations. In future work we plan to extend the models to quantify the time a Remote UE of interest takes to discover all the Relay UEs in proximity and study its impact on the Relay UE selection process. Further extensions of our work will analyze the UE-to-Network Relay discovery protocol under non-ideal channel conditions, considering different propagation environments, and discovery packet recovery based on SINR.

References

[1] S. Lien, C. Chien, G. S. Liu, H. Tsai, R. Li, and Y. J. Wang, “Enhanced LTE Device-to-Device Proximity Services,” *IEEE Communications Magazine*, vol. 54, no. 12, pp. 174–182, Dec 2016.

[2] 3GPP, “Technical Specification Group Core Network and Terminals; Proximity-services (ProSe) User Equipment (UE) to ProSe function protocol aspects; Stage 3 (Release 15) ,” TS 24.334 v15.2.0, 2018.

[3] ——, “Technical Specification Group Services and System Aspects; Proximity-based services (ProSe); Stage 2 (Release 15),” TS 23.303 v15.1.0, 2018.

[4] R. Rouil, F. J. Cintrón, A. Ben Mosbah, and S. Gamboa, “Implementation and Validation of an LTE D2D Model for Ns-3,” in *Proceedings of the 2017 Workshop on Ns-3 (WNS3)*, Jun 2017.

[5] D. Xenakis, M. Kountouris, L. Merakos, N. Passas, and C. Verikoukis, “Performance Analysis of Network-Assisted D2D Discovery in Random Spatial Networks,” *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5695–5707, Aug 2016.

[6] S. Xu and K. S. Kwak, “Network Assisted Device Discovery for D2D Underlying LTE-Advanced Networks,” in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, May 2014.

[7] K. W. Choi and Z. Han, “Device-to-Device Discovery for Proximity-Based Service in LTE-Advanced System,” *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, pp. 55–66, Jan 2015.

[8] H. J. Kang and C. G. Kang, “Performance Analysis of Device-to-Device Discovery with Stochastic Geometry in Non-homogeneous Environment,” in *2014 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct 2014.

[9] H. Bagheri, P. Sartori, V. Desai, B. Classon, M. Al-Shalash, and A. Soong, “Device-to-Device Proximity Discovery for LTE Systems,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*, Jun 2015.

[10] D. Li and Y. Liu, “Performance Analysis for LTE-A Device-to-Device Discovery,” in *2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Aug 2015.

[11] D. Griffith and F. Lyons, “Optimizing the UE Transmission Probability for D2D Direct Discovery,” in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016.

[12] A. Ben Mosbah, D. Griffith, and R. Rouil, “Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery,” in *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, Jan 2018.

[13] D. Griffith, A. Ben Mosbah, and R. Rouil, “Group Discovery Time in Device-to-Device (D2D) Proximity Services (ProSe) Networks,” in *IEEE Conference on Computer Communications (INFOCOM)*, Atlanta, GA, USA, May 2017.

NIST SP 1285
September 2022

Transactions, SMiRT-25
Charlotte, NC, USA, August 4-9, 2019
Division I: Mechanics of Materials

EFFECTS OF ALKALI-SILICA REACTION ON MECHANICAL PROPERTIES AND STRUCTURAL CAPACITIES OF REINFORCED CONCRETE STRUCTURES

Long Phan[1], Fahim Sadek[2], Travis Thonstad[3], H.S. Lew[4], Sorin Marcu[5], Jacob Philip[6]

[1] Leader, Structures Group, NIST, MD, USA (long.phan@nist.gov)
[2] Research Structural Engineer, NIST, MD, USA (fahim.sadek@nist.gov)
[3] Research Structural Engineer, NIST, MD, USA (travis.thonstad@nist.gov)
[4] Senior Research Structural Engineer, NIST, MD, USA (hai.lew@nist.gov)
[5] Engineer Technician, NIST, MD, USA (sorin.marcu@nist.gov)
[6] Senior Geotechnical Engineer, NRC, MD, USA, (Jacob.philip@nrc.gov)

ABSTRACT

This paper describes an ongoing, comprehensive research program being conducted at the National Institute of Standards and Technology (NIST) under the sponsorship of the U.S. Nuclear Regulatory Commission (NRC). The study aims to develop technical basis for evaluating effects of Alkali-Silica Reaction (ASR), which occurs when the high pH concrete pore solution reacts with certain aggregate mineral phases to form expansive ASR gel and create internal expansive forces that cause cracking in concrete, on engineering properties and structural capacities of reinforced concrete structures. Description of this study focusing on experimental evaluation of: (1) the influence of different degrees of steel confinement and ASR expansions (0.1%; 0.3%; 0.5% ultimate expansion) on concrete's mechanical properties; (2) effects of ASR on the bonding between concrete and reinforcements (development and lap splice lengths) and on overall flexural capacities of reinforced concrete beams; and (3) effects of ASR on seismic performance of typical reinforced concrete walls with and without steel confinement in the wall boundary element, will be provided.

INTRODUCTION

Alkali-silica reaction (ASR) has long been recognized as a major cause of concrete internal cracking and deterioration (Stanton 1940; Swenson 1957). This concrete deterioration mechanism begins with reaction between the alkali hydroxides in the cement paste and certain amorphous or micro-crystalline siliceous phases in the aggregates, which produces an alkali-silica gel that forms initially in the partially saturated pore space of the hardened cement paste. The alkali-silica gel is hygroscopic and will continue to absorb moisture in the concrete matrix and expand. This expansion is not reversible and will continue over time as long as moisture is present. The continued expansion of alkali-silica gel creates increasing internal pressure that ultimately leads to internal cracking and degradation of the mechanical properties of concrete (Hansen 1944, Taylor 1990).

Generally, the rate of ASR expansion is relatively slow and is a function of the reactivity of the mineral phases, the alkalinity of the pore solution, and the availability of moisture. Thus, the onset of ASR-induced cracking can take years or decades after construction to occur. However, once occurred, this deterioration at the material level may affect the bonding characteristics between the concrete and the reinforcement and the overall capacity and service life of reinforced concrete structural member or system.

Phan, Long; Sadek, Fahim H.; Thonstad, Travis; Lew, Hai; Marcu, Sorin; Philip, Jacob. "Effects of Alkali-Silica Reaction on Mechanical Properties and Structural Capacities of Reinforced Concrete Structures." Paper presented at Structural Mechanics in Reactor Technology (SMiRT) 25, Charlotte, NC, US. August 04, 2019 - August 09, 2019.

At present, the industry solution is to identify the reactive aggregates and avoid using them through sourcing of materials for construction. This would help with avoiding ASR problem in new construction but does not address the problem in existing structures. Given the current lack of knowledge on ASR effects and standards and codes provisions to account for the effects of ASR on structural capacities, the questions of how to (1) predict the progression of ASR-induced deterioration once occurred and (2) assess the residual material properties and in-situ structural capacity of the affected structures, especially if they are safety-related facilities that are required to maintain a certain safety margin over an extended period of service life, become extremely relevant for certain critical components of the nation's infrastructure (e.g., dams, bridges, and nuclear power plants). Accurate predictions of the progression of ASR and future, residual structural capacities can provide critical support for decision on whether the affected structures can continue safe operation or if it is time for cessation of operations given the increased risk to public safety.

This paper describes work that is part of a comprehensive research program being conducted by the Engineering Laboratory of the National Institute of Standards and Technology (NIST) to study the effects of ASR on the structural performance of nuclear power plant concrete structures. The work is funded by the U.S. Nuclear Regulatory Commission (NRC) under Inter-Agency Agreement NRC-HQ-60-14-I-0004. The objective of this research program is to develop the technical basis for generic regulatory guidance for evaluation of ASR-affected nuclear power plant (NPP) concrete structures throughout its service life. Specifically, the program will develop measurements for evaluation of (1) effects of ASR on structural performance and capability to perform intended function under design basis static and dynamic loads, and (2) characteristics of an asset management program to adequately monitor and manage aging effects of ASR degradation such that intended functions are maintained through the period of extended operation of renewed licenses. The intended outcome is a methodology for determining for an existing ASR-affected structure (1) the in-situ structural capacity to resist design-basis static and dynamic loads and (2) future structural capacity.

NIST EXPERIMENTAL PROGRAM

The NIST experimental program is comprised of two parts. Part 1 focuses on effects of ASR on concrete material properties and performance of reinforced concrete structures subjected to static and pseudo-dynamic loadings. Part 2 examines degradation mechanisms of concrete at micro-structural level. This paper describes Part 1 of the NIST research program focusing on concrete mixture design and curing to achieve certain level of ASR expansion, and the test plan for assessing the effects of ASR on (1) concrete mechanical properties, (2) bonding characteristic and anchorage of steel reinforcement and flexural capacity of reinforced concrete beams, and (3) seismic performance of reinforced concrete walls.

Concrete Mixture Proportioning and Curing

To facilitate examination of the effects of different degrees of ASR expansion on structural capacities of reinforced concrete structural members, four concrete mixtures, namely ASR1 to ASR4, were developed. The first three mixtures, ASR1, ASR2, and ASR3, were designed to have ASR with specific ultimate expansions ε_{ASR} of 0.15%; 0.3%; and 0.5%, respectively. These ultimate expansion levels, from low (ε_{ASR} = 0.15%) to high (ε_{ASR} = 0.5%), were achieved using Type I/II cement with minimum alkali equivalent (Na_2O) content of ≥ 0.87%, alkali reactive coarse aggregate (Placitas, sourced from Albuquerque, New Mexico), and different combinations of reactive (Jobe, sourced from El Paso, TX) and non-reactive (Chaney, sourced from Hagerstown, MD) fine aggregates that meet the grading specification of ASTM C33. The fourth mixture (ASR4) was designed to be non-reactive, utilizing Type I/II cement with low alkali equivalent (Na_2O) content (<0.87%) and non-alkali reactive aggregates, for use in reference specimens for comparison purpose. All coarse aggregates have a maximum size of ¾ inch (20 mm). All four mixtures utilized high range water reducer (BASF MasterGlenium 7710 for the three reactive concrete

Phan, Long; Sadek, Fahim H.; Thonstad, Travis; Lew, Hai; Marcu, Sorin; Philip, Jacob. "Effects of Alkali-Silica Reaction on Mechanical Properties and Structural Capacities of Reinforced Concrete Structures." Paper presented at Structural Mechanics in Reactor Technology (SMiRT) 25, Charlotte, NC, US. August 04, 2019 - August 09, 2019.

mixtures and Sika 2100 for the non-reactive control mixture) that complies with ASTM C494 to aid with workability. Table 1 shows the mixture proportioning for the concretes used in the NIST experimental program.

Table 1: Mixture Proportioning for Concretes in the NIST Experimental Program

Mixture	**Cement Content**			**Coarse Aggregate**			**Fine Aggregate**			**Alkalis ***	**Water**
	Type	**lb/yd^3 (kg/m^3)**	**vol. fraction**	**Type**	**lb/yd^3 (kg/m^3)**	**vol. fraction**	**Type**	**lb/yd^3 (kg/m^3)**	**vol. fraction**	**lb/yd^3 (kg/m^3)**	**Gal (lb/yd^3)**
ASR 1	I/II, high alkali (≥0.87 % Na_2O)	588 (350)	0.111	Placitas	1767 (1050)	0.420	Chaney	1199 (713)	0.274	4.90	35 (294)
ASR 2	"	588 (350)	0.111	"	1767 (1050)	0.420	Jobe	711 (423)	0.274	4.90	35 (294)
							Chaney	480 (285)			
ASR 3	"	588 (350)	0.111	"	1767 (1050)	0.420	Jobe	1185 (705)	0.274	4.90	35 (294)
ASR 4	I/II, low alkali (≤0.87 % Na_2O)	578 (343)	0.115	#57	1805 (1071)	0.399	Washed #33 Sand	1385 (822)	0.295	<2.0	35 (294)

* Total alkali content, including cement alkali + NaOH addition.

Typical curing conditions to accelerate ASR in laboratory conditions involve wetting the top surface prior to form removal and covering the test specimens with wet burlap and plastic sheets after form removal for the first 90 days after casting, at normal ambient temperature and relative humidity, to prevent excessive moisture loss on the specimen's surface layer and minimize potential for drying shrinkage cracks during the hydration process. At 90 days of age, the plastic and burlap covers were removed and the curing temperature and relative humidity in the curing chamber were increased to 75°F (24 ºC) and 75 %. At 180 days of age, the relative humidity was increased again to the range of 95 % to 100 %. The 75°F (24 ºC) and 95 % - 100 % relative humidity curing conditions were then maintained until test time.

Assessing In-Situ Mechanical Properties of ASR-Affected Concrete

This task aims to assess the relationship between ASR-induced expansion and: (1) concrete mechanical properties; (2) surface cracking and expansion, and (3) the influence of hoop reinforcement (i.e., stirrups) on the expansion through measurements conducted on four large block specimens, 3.5 ft × 6 ft × 16 ft (1.07 m × 1.83 m × 4.88 m). Three of the block specimens were made with the ASR1, ASR2, and ASR3 concretes shown in Table 1, and the fourth specimen was made with non-reactive ASR4 concrete and served as the control specimen.

For reinforcement, each block specimen was divided into three regions (Regions 1 to 3, see Figure 1), each region was isolated by a 2-inch (51-mm) thick polystyrene divider to minimize the possible interaction of expansion with the adjoining region and reinforced longitudinally with a combination of #8 and #10 headed bars and transversely with #8 and #10 stirrups and cross ties. The variation in the bar sizes used in each region and inclusion of cross ties in Region 3 were designed to produce different degrees of concrete confinement in the three regions of each block specimen (see Figure 2). Region 2 of each block specimen was not transversely reinforced (no stirrups, reinforcement ratio in the transverse and vertical, or *x* and *y*, directions $\rho_x = \rho_y = 0\%$) and will provide measurements of concrete material properties and cracking characteristics corresponding with the unconfined expansion condition. Region 1 measurements will correspond with intermediate level of expansion confinement ($\rho_x = 0.2\%$), and Region 3 measurements will correspond with a higher level of expansion confinement ($\rho_x = 0.6\%$).

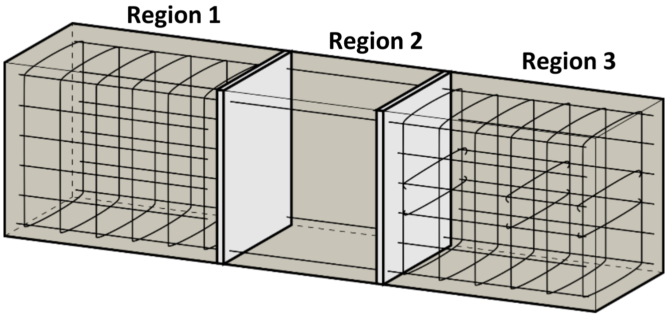


Figure 1. Regions and Reinforcement Scheme of Block Specimen.

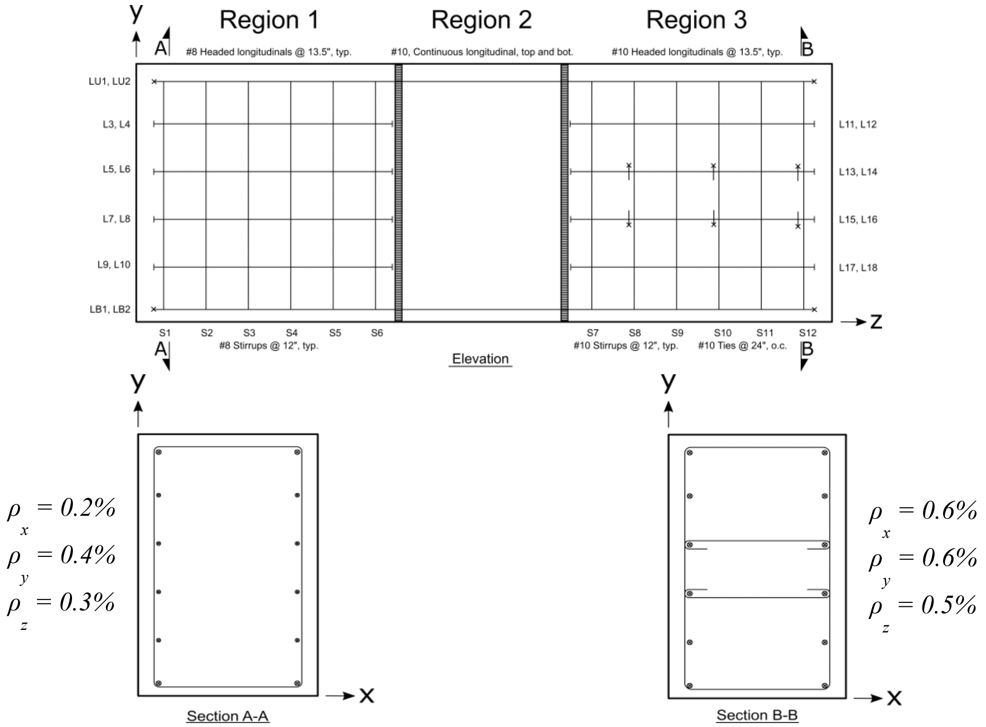


Figure 2. Reinforcement Details of Block Specimen.

All four block specimens were heavily instrumented, each with approximately 164 strain gages attached to the reinforcements, strain transducers to measure tri-directional concrete strain, as well thermo-couples and wireless relative humidity sensors. Figure 3 shows the typical strain gage instrumentation scheme. ASR-induced expansion of concrete, evidenced by increasing strain over time in the reinforcement, are measured and recorded continuously. Figure 4 shows a sample time-history of ASR-induced strain development in the reinforcement in Region 1 of ASR3 block specimen. The time-history provides a correlation between the rates of concrete expansion and the curing conditions and indicates yielding of longitudinal reinforcements shortly before 480 days of curing for this specimen. Correlation of degree of ASR expansion with changes in concrete mechanical properties as a function of time and different levels of confinement is performed through mechanical properties testing of concrete cylinders and cores extracted from different regions of the block specimens. Surface expansion is monitored by measuring changes in vertical and horizontal distances between targets that form square grid (10 ¾ × 10 ¾) on the side of the block specimens using laser tracking device and high-precision calliper.

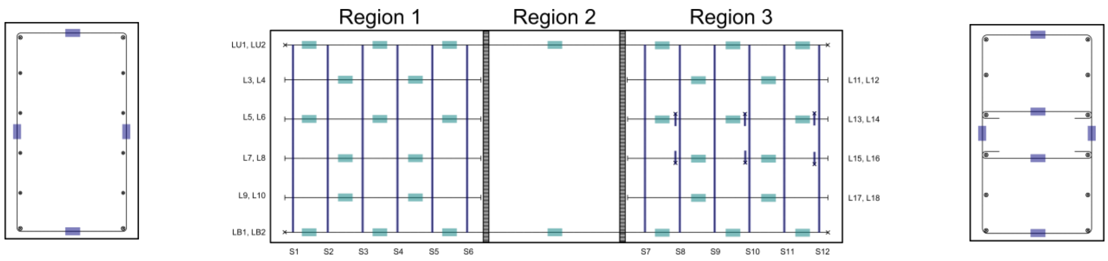


Figure 3. Locations of Strain Gages on the Reinforcements of the Block Specimen.

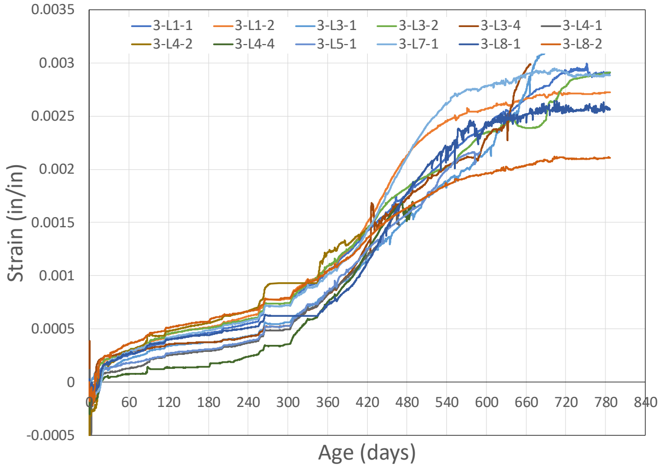


Figure 4. Time-Histories of Strain Development in Region 1 of ASR3 Block Specimen.

Concrete cylinders and cores extracted from different regions of the block specimens are tested periodically for time-histories of concrete compressive strength, splitting tensile strength, and elastic modulus. Data are being processed and compared for evaluation of effects of ASR on concrete mechanical properties. Comparison will also be made with ACI empirical equations to assess applicability of these equations to ASR-affected concrete.

Assessing Bond and Anchorage of Steel Reinforcement

This task aims to examine how ASR influences (1) the bonding characteristic between concrete and steel reinforcement, both for continuous, axially loaded, reinforcing bars and for lap splices in flexural members and (2) the possible loss of rebar anchorage and overall reduction of member's flexural capacities. To facilitate this examination, a series of 19 beam specimens, consisting of 16 reactive beams made of ASR3 concrete and 3 reference, non-reactive beams (also made of ASR3 concrete but with addition of Lithium Nitrate to neutralize the ASR), were constructed and tested under four-point bending. The test series was designed to facilitate examination of the effects of the following three main variables, as well as their interactions, on flexural capacities of reinforced concrete beams: degree of ASR expansion (ε_{ASR}(t)); degree of concrete confinement (K_{tr}/d_b); and ratio of lap splice and development length (l_s/l_d). The test matrix for this test series is shown in Table 2. Beam dimensions and test configuration are shown in Figure 5. Values of $\varepsilon_{ASR}(t)$ shown in Table 2 indicate the nominal expansion level, measured as maximum strain in the reinforcement, relative to the target expansion level $\varepsilon_{ASR}(t)$ = 0.3% for ASR3 concrete when the beam is to be tested (e.g., beam specimen 1, with $\varepsilon_{ASR}(t)$ = 25%, is to be tested when the ASR-induced expansion has resulted in a nominal strain in the reinforcement in the amount of 25% × 0.3% = 0.00075 in/in).

Table 2. Beam Test Matrix

Specimen	**$\varepsilon_{ASR}(t)$** **(% of ASR3 Target Expansion at Test Time)**	**K_{tr} / d_b**	**l_s / l_d**
1	25%	0.5	0.7
2	25%	0.5	1.3
3	25%	1.5	0.7
4	25%	1.5	1.3
5	75%	0.5	0.7
6	75%	0.5	1.3
7	75%	1.5	0.7
8	75%	1.5	1.3
9	50%	1.0	1.0
10	50%	1.0	Continuous Bar
11	Non-Reactive	1.0	1.0
19	100%	1.0	1.0
13	50%	0.0	1.0
14	50%	1.8	1.0
15	50%	1.0	0.5
16	50%	1.0	1.5
17	Non-Reactive	1.0	1.0
18	Non-Reactive	1.0	Continuous Bar
19	100%	1.0	1.0

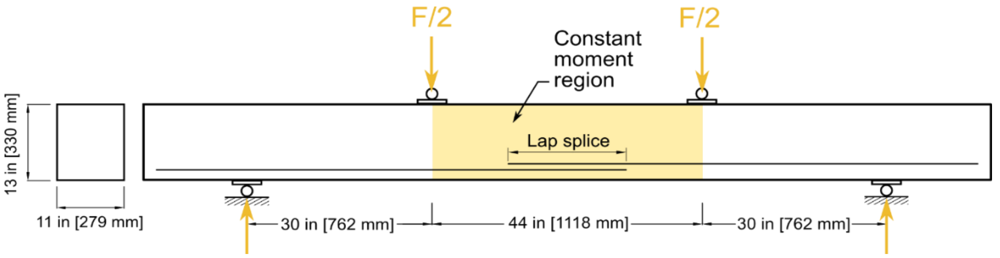


Figure 5. Beam Dimensions and Test Configuration

To date, 16 of 19 beams have been tested. Data analysis is being conducted to examine the effects of the mentioned variables on the bonding characteristic, anchorage strength, and overall flexural beam capacities. Results of the analysis and findings will be presented in a more detailed report to the project sponsor at a later date. Preliminary test results, in the form of maximum vertical load vs. maximum mid-span deflection, are shown in Figure 6. This plot shows the varied modes of failure observed in the 16 tested beams, which ranged from bond failure prior to yielding of the tension reinforcement, yielding of tension reinforcement followed by bond failure, and yielding of tension reinforcement followed by compression failure of concrete in the compression zone.

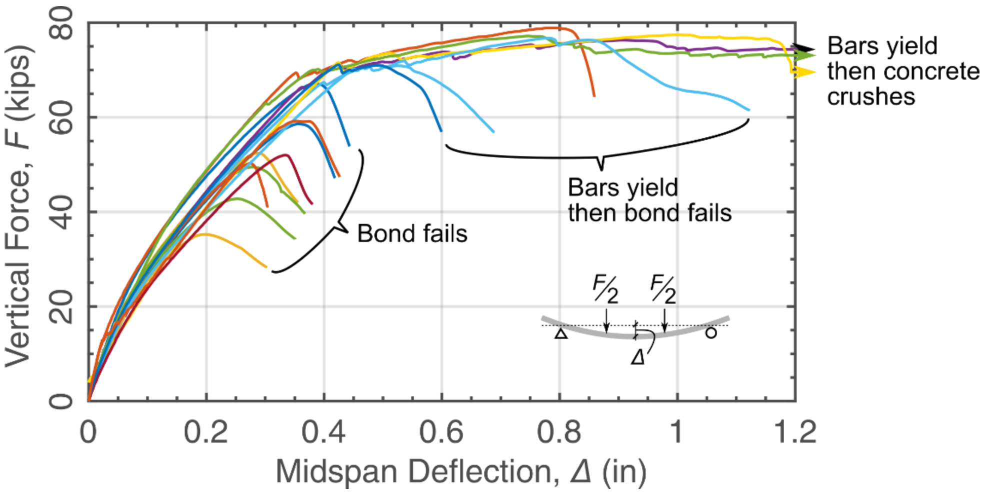


Figure 6. Maximum Load vs. Mid-Span Deflection of Beams

1 kip = 4.45 kN
1 inch = 25.4 mm

Assessing Seismic Performance of ASR-Affected Structural Walls

This task aims to assess the effects of ASR on seismic capacity of typical reinforced concrete walls. Seismic capacities – in terms of lateral stiffness, ultimate strength, ductility, and energy dissipation – of four reinforced concrete wall specimens will be measured through pseudo dynamic cyclic lateral load tests. Three of the wall specimens are made of reactive ASR3 concrete and the fourth non-reactive concrete made of ASR3 concrete modified with Lithium Nitrate (reference specimen). Experimental variables to be studied include (1) degree of ASR expansion $\varepsilon_{ASR}(t)$ and (2) degree concrete confinement, represented by the transverse reinforcement ratio, in the wall's boundary element (ρ_s). Figure 7 shows the dimensions and typical reinforcement details on a cross section of a wall specimen. Table 3 shows the test matrix for this series of wall seismic lateral load test. For testing, the wall specimens are subjected to constant axial compression load to simulate service load while simultaneously subjected to cyclic lateral displacement with increasing magnitude, as shown in Figure 8. At present, the wall specimens have been constructed and are being cured to accelerate ASR expansion for testing.

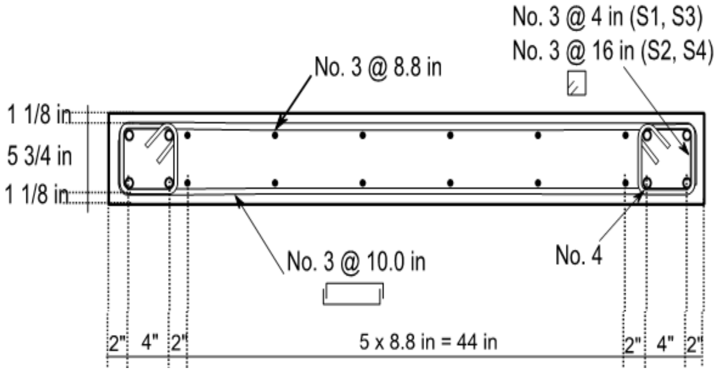


Figure 7. Cross-Sectional Dimensions and Reinforcement Details of Wall Specimen

1 inch = 25.4 mm

Table 3. Wall Lateral Load Test Matrix

Wall Specimen	**$\varepsilon_{ASR}(t)$ (% of ASR3 Ultimate Expansion at Test Time)**	**ρ_s**
1	50%	0%
2	100%	2%
3	100%	0%
4	Non-Reactive	2%

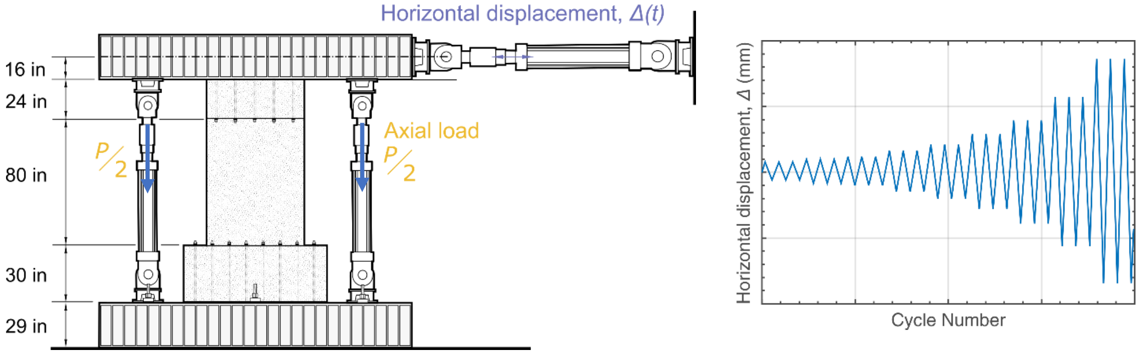

Horizontal displacement, Δ(t)
16 in
24 in
80 in
30 in
29 in
P/2
Axial load
P/2
Horizontal displacement, Δ (mm)
Cycle Number

Figure 8. Wall Test Set-up and Loading Scheme

Phan, Long; Sadek, Fahim H.; Thonstad, Travis; Lew, Hai; Marcu, Sorin; Philip, Jacob. "Effects of Alkali-Silica Reaction on Mechanical Properties and Structural Capacities of Reinforced Concrete Structures." Paper presented at Structural Mechanics in Reactor Technology (SMiRT) 25, Charlotte, NC, US. August 04, 2019 - August 09, 2019.

SUMMARY

An experimental plan to measure and quantify the effects of ASR on concrete material properties and structural capacities of reinforced concrete beam and walls, being conducted at NIST under the sponsorship of the NRC, is described. Measurements obtained from this study will be used to develop a methodology for determining the in-situ and future structural capacity of ASR-affected structure and will form the technical basis underpinning possible generic regulatory guidance for evaluation of ASR-affected nuclear power plant (NPP) concrete structures throughout its service life. Discussions of test results and findings will be provided in detail in subsequent joint NIST/NRC publications.

DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

ASTM C33. 2018. *Standard Specification for Concrete Aggregates.* ASTM International. West Conshocken, PA.

ASTM C494. 2017. *Standard Specification for Chemical Admixtures for Concrete.* ASTM International. West Conshocken, PA

Stanton, T.E. 1940. *"Expansion of concrete through reaction between cement and aggregate."* Proceedings of the American Society of Civil Engineers, 66(10): 1781-1811.

Swenson, E.G. 1957a. *"A reactive aggregate undetected by ASTM tests."* Proceedings of American Society for Testing and Materials, 57, 48-51.

Swenson, E.G. 1957b. "Cement aggregate reaction in concrete of a Canadian bridge." ASTM Proceedings, 57 1043–1056.

LABEL-FREE SURFACE ACOUSTIC WAVE-BASED EMBEDDED FLOW SENSOR

Aurore Quelennec, Jason J. Gorman, and Darwin R. Reyes*
National Institute of Standards and Technology (NIST), Gaithersburg, USA

ABSTRACT

This paper presents a label-free flow sensor embedded in a microfluidic system. This sensor is based on surface acoustic waves, where the acoustic intensity is dependent on the flow rate of the propagating medium. The range of flow rates studied was between 10 µL/min to 1 mL/min, with a sensitivity of 7 µV/(µL/min). Different to readily available sensors that need tracers (*e.g.*, particles, fluorophores) or the use of temperature distribution to measure flow, this sensor requires no external optical components, and can be used with any type of liquid at a broad range of temperatures and liquid conditions.

KEYWORDS: Flow Metrology, Electroacoustics, SAW Sensor

INTRODUCTION

Interdigitated (IDT) electrodes can generate surface acoustic waves (SAW) within a piezoelectric material. The propagation medium interacts with the SAW altering the SAW's acoustic frequency and intensity. This interaction makes SAW transducers useful as both actuators [1] and sensors [2]. SAW have previously been used in microfluidic systems to measure the flow rate [3] of the propagating medium. However, current SAW flow sensors used in microfluidic systems, based on thermal, doppler shift or time-of-flight measurements, are not label-free sensors. In this work, we report a new method for flow measurement over the range of 10 µL/min to 1 mL/min in a microfluidic channel. The technique presented here relates to the ability of the fluid to absorb the acoustic waves as a function of the flow rate.

EXPERIMENTAL

A complete description of the fabrication process can be found in [4]. Three pairs of IDT electrodes (Figure 1) are fabricated using a lift-off process and an e-beam evaporator on a 128 Y-cut lithium niobate (LN) wafer. The electrode is composed of 90 nm of gold on 10 nm of titanium. A 100 nm thick silicon oxide layer is sputtered on the chip. Figure 1A shows the IDT electrode E1 has a single-phase unidirectional transducer (SPUDT) structure, while E2 has a standard IDT structure. Each transducer studied (E1 and E2) has 32 pairs of fingers, which are 3 mm long with a pitch (p) of 80 µm. The 80 µm deep microchannel in polydimethylsiloxane (PDMS) is fabricated using soft lithography techniques.

Our assumption was that the acoustic intensity depends on the flow speed (v_F). E1 is excited with a 10 V peak-to-peak sine wave (V_{in} in Figure 2) at the acoustic resonance frequency. E2 is connected to an oscilloscope to measure the amplitude (V_p) of the received wave at E2 for different v_F. Deionized water is injected in the microchannel using a syringe pump. The flow rate is controlled from 0 µL/min to 1,000 µL/min.

RESULTS AND DISCUSION

As shown in Figure 2 (insets), the peak amplitude increases with an increase in v_F, but the received signal has a low Signal/Noise ratio. We calculated that the average error in the measurement is 1 mV. To extract the sensor response to v_F, the difference between the received signal at different v_F and the one at no flow is calculated (noted as $V_p - V_p(0)$). In Figure 3, the amplitude variation, $V_p - V_p(0)$, of the 1st, 2nd and 5th peaks, is represented as a function of flow, and a linear fit was obtained for each of them. The first peak is not sensitive to v_F, while the others have an average sensitivity of 7 µV/(µL/min). The mean-squared error between the measured value and the linear estimation is 1 mV. These results confirm that the acoustic intensity depends on v_F.

CONCLUSION

This work shows that the acoustic intensity of a SAW is dependent on the flow rate of the propagating medium. This technique represents a step forward towards the development of a calibration-free flow device. This system could, ultimately, become a standard reference device to be used in any microfluidic system and application.

Quelennec, Aurore; Gorman, Jason J.; Reyes-Hernandez, Darwin. "Label-free surface acoustic wave-based embedded flow sensor." Paper presented at The 23rd International Conference on Miniaturized Systems for Chemistry and Life Sciences (µTAS 2019), Basel, CH. October 27, 2019 - October 31, 2019.

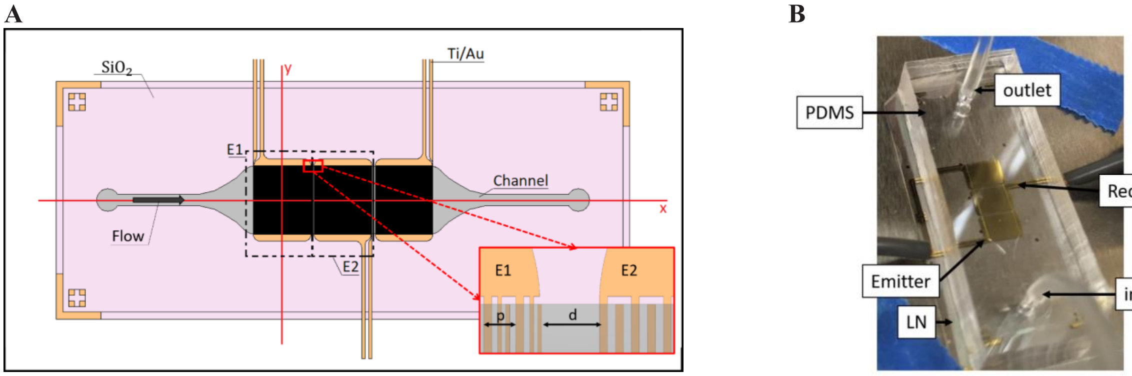
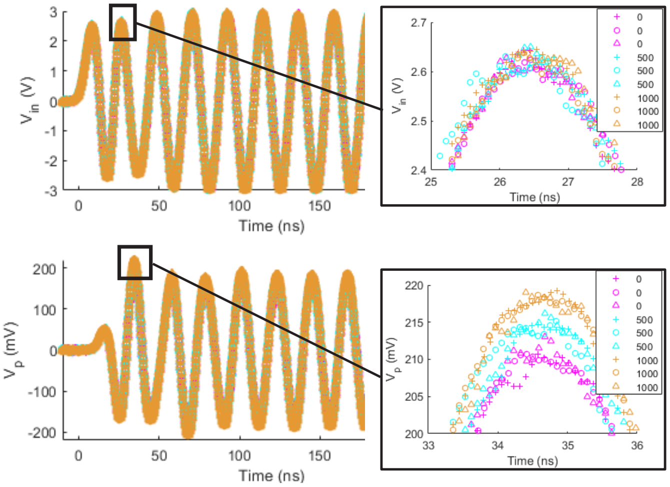


Figure 1: Sensor structure. (A) Top view sketch of the lithium niobate (LN) chip, with 3 pairs of IDT electrodes (in orange), coated with a SiO2 layer (in pink), and the imprint of the microfluidic channel (in grey). A close view of the pairs of IDT electrodes E1 and E2. (B) Image of the chip with the embedded flow sensor under the probe station, where E1 emits and E2 receives.



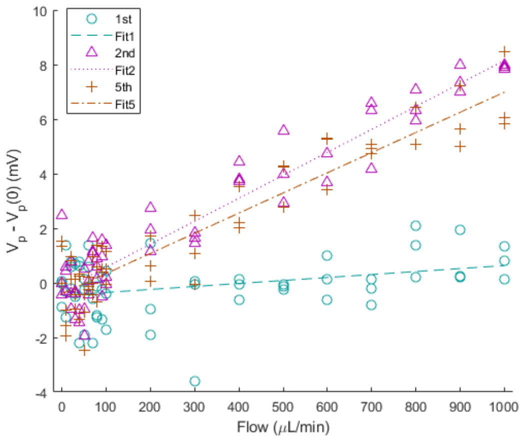


Figure 2: Emitted (V_{in}) and received (V_p) signals at different flow rates (0, 500 and 1000 µL/min). A closer view at each peak shows an increase in peak amplitude with an increase in flow.

Figure 3: Evolution of $V_p - V_p(0)$ of 1st, 2nd and 5th peaks. The 2nd and 5th peaks present a linear response dependent of flow rates with a sensitivity of 7 µV/(µL/min) and an error of 1 mV.

ACKNOWLEDGEMENTS

A.Q., J.J.G. and D.R.R. were supported by the NIST on a Chip Initiative. This research was performed in part at the NIST Center for Nanoscale Science and Technology.

REFERENCES

[1] David B. Go et al., *Anal. Methods*, 2017,**9**, 4112-4134; doi:10.1039/C7AY00690J
[2] Leslie Y. Yeo and James R. Friend, *Annu. Rev. Fluid Mech.,* 2014 46:1, 379-406; doi:10.1146/annurev-fluid-010313-141418
[3] J.T.W. Kuo, L. Yu and E. Meng, *Micromachines* **2012**, *3*(3), 550-573; doi:10.3390/mi3030550
[4] Nathan D. Orloff et al., *Biomicrofluidics,* 2011 5:4; doi:10.1063/1.3661129

CONTACT: Darwin Reyes; phone: +1-301-975-5466; darwin.reyes@nist.gov

Quelennec, Aurore; Gorman, Jason J.; Reyes-Hernandez, Darwin. "Label-free surface acoustic wave-based embedded flow sensor." Paper presented at The 23rd International Conference on Miniaturized Systems for Chemistry and Life Sciences (µTAS 2019), Basel, CH. October 27, 2019 - October 31, 2019.

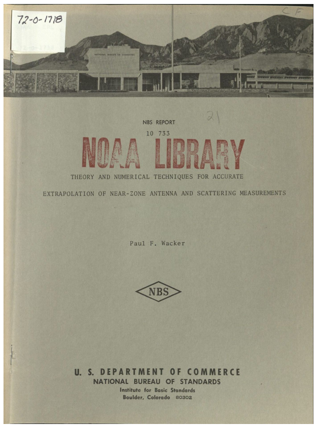
On Wacker's Essential Equation in the Extrapolation Measurement Technique

Alex J. Yuffa
National Institute of Standards and Technology
Boulder, CO 80305, USA
alex.yuffa@nist.gov

***Abstract*—The generalized three-antenna method is a standard method for measuring on-axis gain and polarization of an antenna without a priori knowledge. The cornerstone of the method is the use of the extrapolation technique and the key relationship in the extrapolation technique is Wacker's equation. This equation expresses the received signal as a function of the separation distance between any two antennas. The derivation of Wacker's equation is not readily available in the literature. In this paper, we provide a streamlined derivation of Wacker's equation and address some of the common misconceptions associated with it.**

I. Introduction

The extrapolation technique is commonly used to measure on-axis gain and polarization of an antenna. This technique was introduced almost five decades ago by Newell, Baird, and Wacker in their seminal paper [1]. The technique is based on an equation that represents the received signal as a function of the separation distance between any two antennas. This equation, see equation (26) in [1], is equally accurate when the antennas are close together or far apart because it accounts for the multiple scattering effects between the two antennas. In the seminal paper, the derivation of the equation is attributed to an unpublished technical report by Wacker. Two years before the seminal paper was published Newell and Kerns [2] attributed the derivation of the equation to another of Wacker's unpublished works with an almost identical title. A decade later Kerns, in his classic monograph [3, p. 148], attributes the derivation of the equation to the unpublished report by Wacker. Furthermore, Kerns cites private communication with Yaghjian (presumably Arthur D. Yaghjian) and states that Yaghjian derived an almost identical equation to Wacker's (26) in [1] via a different method [3, p.148]. At this point the referencing of unpublished literature ends and Kerns gives a brute-force derivation of the equation under a greatly simplifying assumption of no multiple scattering [3, pp. 147–159]. It is not clear why the authors cited the report as unpublished as it was *published* almost a year before the seminal paper was submitted for publication [4]. The front page of the report is shown in Fig. 1 and may be obtain from the National Oceanic and Atmospheric Administration (NOAA) library in Boulder, Colorado.


72-0-1718

NBS REPORT
10 733

NOAA LIBRARY

THEORY AND NUMERICAL TECHNIQUES FOR ACCURATE
EXTRAPOLATION OF NEAR-ZONE ANTENNA AND SCATTERING MEASUREMENTS

Paul F. Wacker

NBS

U. S. DEPARTMENT OF COMMERCE
NATIONAL BUREAU OF STANDARDS
Institute for Basic Standards
Boulder, Colorado 80302

Fig. 1. (Color online) The cover of the allusive Wacker's report [4] is shown.

Perhaps because of the perceived lack of availability of Wacker's report, we have encountered a number of practitioners with a misleading interpretation of Wacker's equation. One misconception is that the distance between the two antennas should be measured from the phase centers of the antennas. Another misconception is that Wacker's equation only *approximately* accounts for the multiple scattering effects. This misconception seems to be caused by the input reflection coefficient of the probe antenna in *free space*. To eliminate some of these misconceptions, we present an abridged derivation of Wacker's equation with the aim of pedagogical clarity.

II. Orders-of-Scattering Interpretation

Consider two on-axis antennas separated by free space as shown in Fig. 2. Without lost of generality, we let the antenna under test (AUT) be the antenna on the left-hand side (LHS) in Fig. 2 and the probe antenna to be on the right-hand side (RHS). The scattering-matrix for each antenna is defined by

$$\begin{bmatrix} b_1^\alpha \\ b_2^\alpha \end{bmatrix} = \begin{bmatrix} S_{11}^\alpha & S_{12}^\alpha \\ S_{21}^\alpha & S_{22}^\alpha \end{bmatrix} \begin{bmatrix} a_1^\alpha \\ a_2^\alpha \end{bmatrix}, \tag{1a}$$

where $\alpha = \ell$ for the LHS antenna and $\alpha = \mathrm{r}$ for the RHS antenna. We can also define a scattering-matrix for the space

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

Yuffa, Alex. "On Wacker's Essential Equation in the Extrapolation Measurement Technique." Paper presented at Annual Symposium of the Antenna Measurement Techniques Association (AMTA), San Diego, CA, US. October 06, 2019 - October 11, 2019.

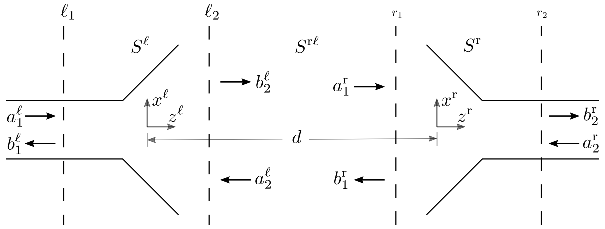


Fig. 2. Two on-axis antennas separated by free space are shown. The terminal surfaces are represented by the labeled dashed lines. The complex amplitudes of the incident and emergent modes are denoted by $a_1^\ell, a_2^{\mathrm{r}}$ and $b_1^\ell, b_2^{\mathrm{r}}$, respectively.

between the antennas via

$$\begin{bmatrix} a_1^{\mathrm{r}} \\ a_2^\ell \end{bmatrix} = \begin{bmatrix} 0 & S_{12}^{\mathrm{r}\ell} \\ S_{21}^{\mathrm{r}\ell} & 0 \end{bmatrix} \begin{bmatrix} b_1^{\mathrm{r}} \\ b_2^\ell \end{bmatrix}, \tag{1b}$$

where the diagonal elements vanish because there is no coupling between the incident and emergent waves at the r_1 terminal surface and the ℓ_2 terminal surface. The scattering-matrix for the whole system is defined by

$$\begin{bmatrix} b_1^\ell \\ b_2^{\mathrm{r}} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} a_1^\ell \\ a_2^{\mathrm{r}} \end{bmatrix} \tag{2}$$

and can be expressed in terms of $S^\ell, S^{\mathrm{r}\ell}, S^{\mathrm{r}}$ by eliminating a_2^ℓ, b_2^ℓ and $a_1^{\mathrm{r}}, b_1^{\mathrm{r}}$ from (1) and comparing the result to (2) to obtain

$$S_{11} = S_{11}^\ell + S_{12}^\ell R^{\mathrm{r}} \left[1 - S_{22}^\ell R^{\mathrm{r}}\right]^{-1} S_{21}^\ell, \tag{3a}$$

$$S_{12} = S_{12}^\ell S_{21}^{\mathrm{r}\ell} \left[1 - S_{11}^{\mathrm{r}} R^\ell\right]^{-1} S_{12}^{\mathrm{r}}, \tag{3b}$$

$$S_{21} = S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} \left[1 - S_{22}^\ell R^{\mathrm{r}}\right]^{-1} S_{21}^\ell, \tag{3c}$$

$$S_{22} = S_{22}^{\mathrm{r}} + S_{21}^{\mathrm{r}} R^\ell \left[1 - S_{11}^{\mathrm{r}} R^\ell\right]^{-1} S_{12}^{\mathrm{r}}, \tag{3d}$$

where

$$R^{\mathrm{r}} = S_{21}^{\mathrm{r}\ell} S_{11}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} \qquad \text{and} \qquad R^\ell = S_{12}^{\mathrm{r}\ell} S_{22}^\ell S_{21}^{\mathrm{r}\ell} \tag{4}$$

In (4), R^{r} (R^ℓ) is the reflection coefficient of the RHS (LHS) antenna as seen by the LHS (RHS) antenna.

By the nth orders-of-scattering approximation we mean an expansion of (3) in "powers" of R^{r} or R^ℓ up to and including order n [5]. It is important to realize that although we have treated the elements of the scattering-matrices as scalar numerical entities they are not. In general, the elements of the scattering-matrices are integral operators, and thus (2) contains not algebraic equations but rather integral equations in disguise. Therefore, the operators in the square brackets in (3) should be expanded in the Neumann series [6, §3.2],

$$[1 - G]^{-1} = 1 + G + GG + \cdots = \sum_{n=0}^{\infty} G^n, \quad ||G|| < 1, \tag{5}$$

rather than the Taylor series. In other words, the G^n term in (5) should be interpreted as the nth iterated kernel and 1 should be interpreted as the identity operator.

To gain some physical insight into (5) let's consider the 2nd orders-of-scattering approximation of S_{21} and S_{22}. In other words, expanding the square bracket terms in S_{21} and S_{22} in the Neumann series yields

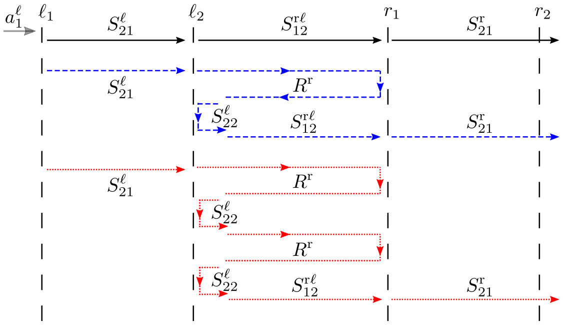


Fig. 3. (Color online) The 2nd orders-of-scattering approximation of $S_{21}^{(2)}$ is schematically illustrated.

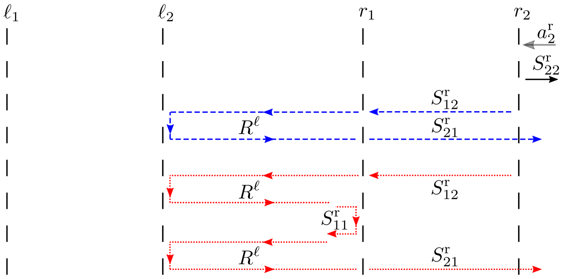


Fig. 4. (Color online) The 2nd orders-of-scattering approximation of $S_{22}^{(2)}$ is schematically illustrated.

$$S_{21}^{(2)} = S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} S_{21}^\ell + S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} \left(S_{22}^\ell R^{\mathrm{r}}\right) S_{21}^\ell + S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} \left(S_{22}^\ell R^{\mathrm{r}}\right)^2 S_{21}^\ell, \tag{6a}$$

and

$$S_{22}^{(2)} = S_{22}^{\mathrm{r}} + S_{21}^{\mathrm{r}} R^\ell S_{12}^{\mathrm{r}} + S_{21}^{\mathrm{r}} R^\ell \left(S_{11}^{\mathrm{r}} R^\ell\right) S_{12}^{\mathrm{r}}, \tag{6b}$$

where the superscript (2) denotes the order of the approximation. In Fig. 3, the first, second, and third terms on the RHS of (6a) are schematically shown by the solid, dashed, and dotted lines, respectively. From Fig. 3, we see that the first term corresponds to the direct transmission of a_1^ℓ and the second term includes the reflections of the signal by the probe and the AUT. A similar interpretation may be constructed for (6b), see Fig. 4. It is important to note that the first term on the RHS of (6b) corresponds to the direct reflection of a_2^{r} by the RHS antenna (probe) as seen from its feed. In other words, this is the reflection coefficient that one would measure in the absence of the LHS antenna (AUT).

III. Wacker's Equation

In the previous section, we provided a physical insight into the mutual coupling between the two antennas. This was done in a general and abstract manner by formal manipulation of the scattering-matrices. In this section, we will remove the abstraction layer by constructing an explicit form of the S-matrix.

If we assume the probe antenna is connected to a load with the reflection coefficient Γ^{r}, then $a_2^{\mathrm{r}} = \Gamma^{\mathrm{r}} b_2^{\mathrm{r}}$ and from (2) we have

$$b_2^{\mathrm{r}} = \left[1 - S_{22}\Gamma^{\mathrm{r}}\right]^{-1} S_{21} a_1^{\ell}. \tag{7}$$

Notice that the square bracket on the RHS of (7) contains the integral operator S_{22} and *not* the reflection coefficient of the probe S_{22}^{r}.

A. Zeroth Order-of-Scattering

To obtain the zeroth order-of-scattering approximation of (7) we substitute the first term on the RHS of (6) into (7) to obtain

$$b_2^{\mathrm{r}(0)} = \left[1 - S_{22}^{\mathrm{r}}\Gamma^{\mathrm{r}}\right]^{-1} S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} S_{21}^{\ell} a_1^{\ell}. \tag{8}$$

In (8), the term inside the square bracket is a complex *scalar* quantity, and thus we only need to determine the explicit form of

$$T \equiv S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell} S_{21}^{\ell} a_1^{\ell}. \tag{9}$$

The key to obtaining the explicit form of T is to recognize that it may be written as [4, pp.12–13]

$$T = \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} a_1^{\ell} f_1(k_x) f_2(k_y) f_3(k_z) \frac{\mathrm{e}^{+\mathrm{i}k_z d}}{k_z}\, \mathrm{d}k_x \mathrm{d}k_y, \tag{10a}$$

where f_1, f_2, and f_3 are *entire* functions and wave vector $\boldsymbol{k} = k_x \hat{\boldsymbol{e}}_{x^\ell} + k_y \hat{\boldsymbol{e}}_{y^\ell} + k_z \hat{\boldsymbol{e}}_{z^\ell}$ is such that

$$k_z = \begin{cases} +\sqrt{k^2 - \kappa^2} & \text{for} \quad \kappa < k \\ +\mathrm{i}\sqrt{\kappa^2 - k^2} & \text{for} \quad \kappa > k \end{cases}, \tag{10b}$$

with $\kappa^2 = k_x^2 + k_y^2$, see Fig. 2. Recall that for an entire function the principal part of the Laurent series vanishes [7, p.17], and thus we have

$$f_1 = \sum_{m=0}^{\infty} A_m^x k_x^m, \quad f_2 = \sum_{n=0}^{\infty} A_n^y k_y^n, \quad f_3 = \sum_{p=0}^{\infty} A_p^z k_z^p. \tag{11}$$

Substituting (11) into (10a), and noting that the odd powers of k_x and k_y integrate to zero, we obtain

$$T = a_1^{\ell} \sum_{mnp} A_m^x A_n^y A_p^z \times \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty} k_x^{2m} k_y^{2n} k_z^p \frac{\mathrm{e}^{+\mathrm{i}k_z d}}{k_z}\, \mathrm{d}k_x \mathrm{d}k_y, \tag{12}$$

where the triple sum is over all non-negative integers. Converting the double integral in (12) to polar coordinates via $k_x = \kappa\cos\theta$ and $k_y = \kappa\sin\theta$, then using the orthogonal property of cosines and sines yields

$$T = a_1^{\ell} \sum_{n=0}^{\infty}\sum_{p=0}^{\infty} A_n^x A_n^y A_p^z \int_0^{\infty} \kappa^{4n} k_z^p \frac{\mathrm{e}^{+\mathrm{i}k_z d}}{k_z} \kappa\, \mathrm{d}\kappa. \tag{13}$$

Changing the integration variable in (13) from κ to k_z by recalling that κ^2 is a function of k_z^2, see (10b), yields

$$T = -a_1^{\ell} \sum_{n=0}^{\infty}\sum_{p=0}^{\infty} A_n^x A_n^y A_p^z \times \int_k^{\mathrm{i}\infty} \left(k^2 - k_z^2\right)^{2n} k_z^p \mathrm{e}^{+\mathrm{i}k_z d}\, \mathrm{d}k_z. \tag{14}$$

If we parameterize the line integral in the complex k_z-plane in (14) via $k_z = k + \mathrm{i}t$, then the product of the two infinite sums in (14) can be written as

$$\sum_{n=0}^{\infty}\sum_{p=0}^{\infty} A_n^x A_n^y A_p^z \left(k^2 - k_z^2\right)^{2n} k_z^p = \sum_{q=0}^{\infty} B_q t^q, \tag{15}$$

where the unknown B_q coefficients depend on the A_n^x, A_n^y, A_p^z coefficients and the wavenumber k. Substituting (15) into (14) and changing the integration variable from k_z to t yields

$$T = -\mathrm{i}a_1^{\ell} \sum_{q=0}^{\infty} B_q \mathrm{e}^{+\mathrm{i}kd} \int_0^{\infty} t^q \mathrm{e}^{-td}\, \mathrm{d}t. \tag{16}$$

Recognizing the integral in (16) as the Laplace transform of t^q [8, §17.13] we immediately obtain

$$T = a_1^{\ell} \frac{\mathrm{e}^{+\mathrm{i}kd}}{kd} \sum_{n=0}^{\infty} \frac{A_n}{(kd)^n}, \tag{17}$$

where the unknown A_n coefficients depend on the B_q coefficients. The form of (17) is not surprising because according to the Wilcox expansion theorem [9] any electromagnetic wave produced by a *finite* charge distribution has an expansion of the form

$$\frac{\mathrm{e}^{+\mathrm{i}kr}}{kr} \sum_{n=0}^{\infty} \frac{\boldsymbol{A}_n(\theta,\phi)}{(kr)^n}, \qquad r > r_c, \tag{18}$$

where (r, θ, ϕ) are the usual spherical coordinates and r_c is the radius of the smallest circumscribing sphere containing the finite charge distribution. Finally, after substituting (17) into (8) we obtain the zeroth order-of-scattering approximation of the signal measured by the probe; namely,

$$b_2^{\mathrm{r}(0)} = a_1^{\ell} \left[1 - S_{22}^{\mathrm{r}}\Gamma^{\mathrm{r}}\right]^{-1} \frac{\mathrm{e}^{+\mathrm{i}kd}}{kd} \sum_{n=0}^{\infty} \frac{A_n}{(kd)^n}. \tag{19}$$

From the derivation of (17) we see that the A_n coefficients in (19) are *independent* of S_{22}^{r} and Γ^{r}. This observation is obvious in the current context but it will become camouflaged when we consider all orders-of-scattering.

B. All Orders-of-Scattering

The first orders-of-scattering approximation of S_{21} includes the first two terms on the RHS of (6a); namely, the zeroth order-of-scattering, which we computed in Section III-A, and the $S_{21}^{\mathrm{r}} S_{12}^{\mathrm{r}\ell}\left(S_{22}^{\ell} R^{\mathrm{r}}\right) S_{21}^{\ell}$ term. Similar to the computation of the zeroth order-of-scattering approximation, the key to obtaining the explicit form of $\mathbb{T} = S^{\mathrm{r}}_{21} S^{\mathrm{r}\ell}_{12} \left(S^{\ell}_{22} R^{\mathrm{r}}\right) S^{\ell}_{21} a^{\ell}_{1}$ is to recognize that it may be written as [4, p.24]

$$\mathbb{T} = a_1^{\ell} \iint\limits_{-\infty}^{\infty} \mathrm{d}k_x \mathrm{d}k_y \iint\limits_{-\infty}^{\infty} \mathrm{d}k'_x \mathrm{d}k'_y \iint\limits_{-\infty}^{\infty} \mathrm{d}k''_x \mathrm{d}k''_y$$

$$f(k_x, k_y, k'_x, k'_y, k''_x, k''_y) \frac{\mathrm{e}^{\mathrm{i}(k_z + k'_z + k''_z)d}}{k_z + k'_z + k''_z}, \quad (20)$$

where the function f is an entire function in each variable. The integrals in (20) may be evaluated following the procedure of Section III-A to obtain

$$\mathbb{T} = a_1^{\ell} \frac{\mathrm{e}^{3\mathrm{i}kd}}{(kd)^3} \sum_{n=0}^{\infty} \frac{A'_n}{(kd)^n}. \quad (21)$$

In general, we can use the above method for all orders-of-scattering. Therefore, we can obtain *exact* explicit forms of S_{21} and S_{22}. These forms are given by [4, pp.25–27],

$$S_{21} = \frac{\mathrm{e}^{\mathrm{i}kd}}{kd} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{2m\mathrm{i}kd}}{(kd)^{2m}} \frac{A_{mn}}{(kd)^n}. \quad (22a)$$

and

$$S_{22} = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{2m\mathrm{i}kd}}{(kd)^{2m}} \frac{B_{mn}}{(kd)^n}, \quad (22b)$$

where $B_{0n} = S^{\mathrm{r}}_{22} \delta_{0n}$ and δ_{0n} is the Kronecker delta function. In (22b), $B_{0n} = S^{\mathrm{r}}_{22} \delta_{0n}$ because (22b) must agree with the zeroth order-of-scattering approximation, i.e., $S^{(0)}_{22} = S^{\mathrm{r}}_{22}$.

To obtain Wacker's equation, we use the Neumann series to expand the square bracket term in (7) and then substitute (22) into the resultant, i.e.,

$$\frac{b^{\mathrm{r}}_2}{a^{\ell}_1} = \frac{\mathrm{e}^{\mathrm{i}kd}}{kd}$$

$$\times \sum_{p=0}^{\infty} \left[\Gamma^{\mathrm{r}} \sum_{mn} \frac{\mathrm{e}^{2m\mathrm{i}kd}}{(kd)^{2m}} \frac{B_{mn}}{(kd)^n} \right]^p \sum_{m'n'} \frac{\mathrm{e}^{2m'\mathrm{i}kd}}{(kd)^{2m'}} \frac{A_{m'n'}}{(kd)^{n'}} \quad (23)$$

Wacker's equation (23) may be written in a more traditional form by multiplying out the sums in (23) to obtain

$$\frac{b^{\mathrm{r}}_2}{a^{\ell}_1} = \frac{\mathrm{e}^{\mathrm{i}kd}}{kd} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{2m\mathrm{i}kd}}{(kd)^{2m}} \frac{C_{mn}}{(kd)^n}, \quad (24)$$

where, in general, the C_{mn} coefficients depend on the A_{mn}, B_{mn} coefficients and Γ^{r}. We can simplify (24) further by noting that for $m = 0$ it must agree with the zeroth order-of-scattering approximation given by (19); thus, we have

$$\frac{b^{\mathrm{r}}_2}{a^{\ell}_1} = \frac{1}{1 - S^{\mathrm{r}}_{22}\Gamma^{\mathrm{r}}} \frac{\mathrm{e}^{\mathrm{i}kd}}{kd} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{2m\mathrm{i}kd}}{(kd)^{2m}} \frac{F_{mn}}{(kd)^n}, \quad (25)$$

where F_{00} and only F_{00} is *independent* of S^{r}_{22} and Γ^{r} (see the end of Section III-A).

IV. Discussion

Wacker's equation is given by (25) and it accounts for *all* multiple scattering effects between the two antennas. Unfortunately, it can also be improperly derived leading to misconceptions. To see this, substitute the *exact* form of S_{21} given by (22a) into the zeroth order-of-scattering approximation (8) to obtain

$$\frac{1}{1 - S^{\mathrm{r}}_{22}\Gamma^{\mathrm{r}}} \frac{\mathrm{e}^{\mathrm{i}kd}}{kd} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{2m\mathrm{i}kd}}{(kd)^{2m}} \frac{A_{mn}}{(kd)^n}. \quad (26)$$

This improper derivation suggests that Wacker's equation accounts for the multiple scattering effects contained in S_{22} only to the zeroth order-of-scattering. Of course, this conclusion is incorrect as we have shown in Section III-B.

Another source of confusion comes from the separation distance variable d. In general, d should be measured from behind the face of the AUT to behind the face of the probe as shown in Fig. 2. With this choice, the sums in (25) will usually converge for any d as long as the two antennas are not touching [4, pp.15–22]. A sufficient, but not necessary, condition for the sums in (25) to converge is given by

$$d > r_{\text{AUT}} + r_{\text{probe}}, \quad (27)$$

where r_{AUT} (r_{probe}) is the radius of the smallest circumscribing sphere containing the AUT (probe) [3, Ch. III, §5]. It is important to note that if the sums in (25) converge, then the on-axis gain and polarization of the antenna are *not* affected by the choice of origin of the separation distance. To see this, we shift the coordinate system by d_0 so that the new separation distance is given by

$$d' = d - d_0 \quad (28)$$

and substitute (28) into (25) to obtain

$$\frac{b^{\mathrm{r}}_2}{a^{\ell}_1} = \frac{1}{1 - S^{\mathrm{r}}_{22}\Gamma^{\mathrm{r}}} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{(2m+1)\mathrm{i}kd'}}{(kd')^{2m+1+n}} \frac{F_{mn}\mathrm{e}^{(2m+1)\mathrm{i}kd_0}}{(1+\varepsilon)^{2m+n+1}}, \quad (29)$$

where $\varepsilon = d_0/d'$. Then, expanding $1/(1+\varepsilon)^{2m+n+1}$ in the binomial series and relabeling the expansion coefficients yields

$$\frac{b^{\mathrm{r}}_2}{a^{\ell}_1} = \frac{1}{1 - S^{\mathrm{r}}_{22}\Gamma^{\mathrm{r}}} \frac{\mathrm{e}^{\mathrm{i}kd'}}{kd'} \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{\mathrm{e}^{2m\mathrm{i}kd'}}{(kd')^{2m}} \frac{F'_{mn}}{(kd')^n}, \quad (30)$$

where $F'_{00} = \mathrm{e}^{\mathrm{i}kd_0} F_{00}$. In the far field, the gain and polarization are independent of the $F'_{mn}, m > 0, n > 0$ coefficients. Thus, if we tactfully assume that the different polarization states of the antenna are measured in the same coordinate system, then from $F'_{00} = \mathrm{e}^{\mathrm{i}kd_0} F_{00}$ we see that the gain and polarization of the antenna are unchanged by an absolute phase shift [1, p.427]. In other words, only the relative phase between the different polarization states is of consequence.

V. Conclusions

In this paper, we provided orders-of-scattering interpretation of the scattering-matrix for a two antenna system. We obtained the orders-of-scattering interpretation by first expressing the scattering-matrix of the system in terms of its individual components before formally expanding it in the Neumann series, see Section II. Using the series representation of the scattering-matrix we obtained the explicit forms of its elements and derived Wacker's equation. We showed that Wacker's equation is exact and discussed its convergence properties. Furthermore, we also discussed from where the separation distance between the two antennas should be measured and showed that the choice of the origin does not affect the gain and the polarization of the antenna.

References

[1] A. Newell, R. Baird, and P. Wacker, "Accurate measurement of antenna gain and polarization at reduced distances by an extrapolation technique," *IEEE Transactions on Antennas and Propagation*, vol. 21, no. 4, p. 418–431, Jul. 1973.

[2] A. C. Newell and D. M. Kerns, "Determination of both polarisation and power gain of antennas by a generalised 3-antenna measurement method," *Electronics Letters*, vol. 7, no. 3, p. 68, 1971. [Online]. Available: http://dx.doi.org/10.1049/el:19710047

[3] D. M. Kerns, *Plane-wave scattering-matrix theory of antennas and antenna-antenna interactions*, ser. NBS monograph 162. Washington, DC: U.S. Government Printing Office, 1981.

[4] P. F. Wacker, "Theory and numerical techniques for accurate extrapolation of near-zone antenna and scattering measurements," National Bureau of Standards, U.S. Department of Commerce, NBS Report 10 733, Apr. 1972.

[5] A. J. Yuffa, P. A. Martin, and J. A. Scales, "Scattering from a large cylinder with an eccentrically embedded core: An orders-of-scattering approximation," *J. Quant. Spectrosc. Radiat. Transfer*, vol. 133, p. 520–525, Jan. 2014.

[6] I. Stakgold, *Boundary Value Problems of Mathematical Physics*, ser. Classics in applied mathematics. Society for Industrial and Applied Mathematics, 2000, vol. 1, no. 29.

[7] A. I. Markushevich, *Theory of Functions of a Complex Variable*, 2nd ed., R. A. Silverman, Ed. Chelsea, 1985, vol. 2, three volumes in one.

[8] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed., A. Jeffrey and D. Zwillinger, Eds. Boston: Academic Press, 2007.

[9] C. H. Wilcox, "An expansion theorem for electromagnetic fields," *Communications on Pure and Applied Mathematics*, vol. 9, no. 2, p. 115–134, may 1956.

Verifying the Performance of a Correlation-Based Channel Sounder in the 3.5 GHz Band with a Calibrated Vector Network Analyzer

Jeffrey A. Jargon, Jeanne T. Quimby, Kate A. Remley, Amanda A. Koepke, and Dylan F. Williams
National Institute of Standards and Technology
Boulder, CO 80305 USA
Email: jeffrey.jargon@nist.gov

***Abstract* — We verified the performance of a correlation-based channel sounder in the 3550 MHz to 3650 MHz band with a calibrated vector network analyzer (VNA) by comparing measurements in a stable, coaxial environment at the same reference planes. The purpose of this experiment was to focus on the performance of the channel sounder's hardware, as opposed to antenna effects or channel variations. Two conducted propagation channels were utilized – one consisting of a length of cable and an attenuator to simulate a line-of-sight channel, and another with a pair of splitters joined by cables of different lengths to simulate a multipath environment. Performing repeated measurements and estimating the components of uncertainty due to random effects, we found that the channel sounder and VNA measurements agreed to within 0.25 dB for values of path gain, and the peaks in the power delay profile agreed to within 2 dB.**

***Keywords* — comparison, conducted channel, correlation-based channel sounder, measurement, path gain, power delay profile, vector network analyzer, wireless systems.**

I. Introduction

Channel sounding is a method for evaluating the electromagnetic environment for wireless communications. Radio signals propagate between a transmitter and receiver over multiple paths due to natural terrain, manmade obstacles, and environmental conditions [1]. Channel sounders measure characteristics of a radio propagation channel including path gain, decay time, and angular dispersion. Models developed from these measurements are typically the first step in standardizing new wireless technologies. While many models currently exist, applications are constantly under development that necessitate new and improved channel models. For example, in the 3550 MHz to 3650 MHz ("3.5 GHz") band, rules for spectrum-sharing systems are being developed based on specific channel models [2]. Success of the spectrum-sharing systems will, in part, depend on the accuracy of these models.

Recently, researchers at the U.S. National Institute of Standards and Technology (NIST) and the Institute for Telecommunication Sciences (ITS) collaborated to conduct a series of channel-sounder verifications to identify sources of uncertainty due to systematic and random effects in channel-sounder hardware. Three separate channel sounders operating in the 3.5 GHz frequency band, each having significantly different architectures, were studied [3]. Uncertainties in the channel sounders were studied by comparing coaxial measurements made by these systems to measurements performed on a VNA, which included reference-plane translations for direct comparison and a thorough uncertainty analysis.

Conducted-channel measurements were performed to focus on the performance of the channel sounders' hardware, as opposed to antenna effects or channel variations. Two propagation channels were studied – one consisting of a length of coaxial cable and an attenuator to simulate a line-of-sight channel, and another with a pair of splitters joined by coaxial cables of different lengths to simulate a multipath environment.

Here, we focus on the comparison between the VNA and the NIST correlation-based channel sounder using path gain and power delay profile (PDP) as the metrics. Path gain is a measure of attenuation an electromagnetic field experiences as it propagates through space. We have slightly modified this definition to include propagation through a conducted channel. PDP gives the intensity of a signal received through a multipath channel as a function of time delay, which is the difference in travel time between multipath arrivals [4].

Propagation of the VNA's uncertainties to the metrics of interest (path gain and PDP) were performed using NIST's Microwave Uncertainty Framework software [5]. Furthermore, we shifted the VNA reference planes to align directly with the channel sounder's reference planes using characterized switches that connected both systems to the conducted channels. This scheme, illustrated in Figure 1, enabled direct comparisons. We also studied the variability in measurements with respect to various timescales of relevance to channel measurements, including repeated measurements conducted in rapid succession, measurements made on an hour-scale timeframe, and measurements reproduced over several days.

Below, we describe the correlation-based channel sounder, explain how we derived path gain and PDP from VNA measurements, and provide results of our comparison.

II. Correlation-Based Channel Sounder

The channel sounder is a Pseudo-Noise (PN)-sequence correlation-based system [6]. It consists of a transmitter and receiver synchronized with two rubidium clocks. The synchronized clocks ensure that drift between samples is small enough for accurate resolution of the delay spread and allows for measuring the absolute timing between transmitter and receiver.

Jargon, Jeffrey; Quimby, Jeanne; Remley, Kate; Koepke, Amanda; Williams, Dylan. "Verifying the Performance of a Correlation-Based Channel Sounder in the 3.5 GHz Band with a Calibrated Vector Network Analyzer." Paper presented at Asia-Pacific Microwave Conference (APMC) 2019, Singapore, SG. December 10, 2019 - December 13, 2019.

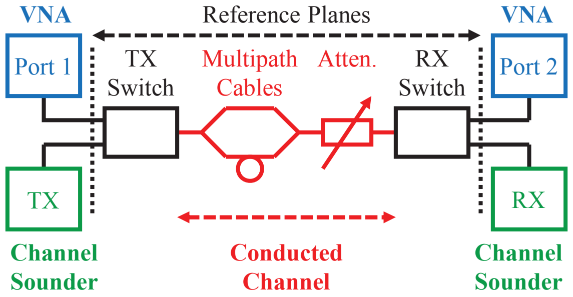


Fig. 1. Measurement setup.

The channel sounder's transmitter contains a vector signal generator that generates a PN code sequence and modulates a binary-phase-shift-keying (BPSK) signal since this type of modulation is robust and widely used in long-distance communications [7]. The transmitter then upconverts the BPSK signal to the 3.5 GHz RF carrier frequency. The bandwidth of the channel sounder corresponds to the bandwidth of the BPSK symbols modulated by the PN sequence. The vector signal generator we used is specified to have a maximum output power of +10 dBm with a -161 dBm/Hz noise floor. The waveform corresponding to each PN sequence is oversampled by four times, providing 8188 samples with a 5 nanosecond/symbol sampling rate. Therefore, a single record of 400 PN sequences (or "code words") has a duration of 16.37 milliseconds.

The channel sounder's transmitter repetitively transmits a PN sequence of digital symbols with a maximum sequence length of 2047 (order of 11). The average power transmitted is maintained through the continuous transmission of the signal. The modulated signal is amplified and filtered to reduce harmonics. The signal is transmitted either through an attenuator for a back-to-back measurement or through the pair of splitters joined by coaxial cables of different lengths.

The channel sounder measures a set of raw complex data, which are corrected with a back-to-back measurement to remove hardware effects. The corrected data are then used to estimate path gain and PDP of the measured channel. Further details of this system can be found in Reference [6].

III. Vector Network Analyzer

Channel sounding can be performed in the frequency domain with a VNA, however, this approach is typically too slow in dynamic environments to capture the time-varying nature of a wireless channel across a wide bandwidth. Furthermore, VNAs are problematic for channels where the transmitter and receiver need to be spaced far apart since the VNA requires that the transmit and receive sides must be tethered together for timing synchronization. Other systems, such as the NIST correlation-based channel sounder, are usually preferred since they can make measurements quickly and are equipped with rubidium clocks to provide untethered synchronization.

In the case of stable conducted channels, VNAs are ideal for verification purposes since they have a high dynamic range, can make extremely wideband channel measurements (dependent on the bandwidth of the VNA), and are capable of traceable uncertainties, as explained below.

In this experiment, we made use of a short-open-load-thru (SOLT) calibration kit with Type-N coaxial connectors. Physical models of the calibration standards were developed and validated using a TRL calibration within the NIST Microwave Uncertainty Framework [8]. This software tool utilizes parallel sensitivity and Monte-Carlo analyses, and enables us to capture and propagate the *S*-parameter measurement uncertainties and statistical correlations between them. By identifying and modeling the physical error mechanisms in the calibration standards, we can determine the statistical correlations among the *S*-parameters. These uncertainties, due to systematic effects, can then be propagated to measurements of devices under test. The uncertainties are propagated to the calculated metrics of interest (path gain and PDP) while maintaining correlated uncertainty mechanisms throughout the process.

PDP and path gain may be calculated from *S*-parameter measurements. Prior to computing these metrics, the VNA hardware settings were chosen in consideration of the channel and channel sounder. The IF bandwidth of the VNA was set to 20 Hz to ensure adequate dynamic range (greater than 110 dB). The VNA frequency range was set to 3.3–3.7 GHz, which was the range used by the correlation-based channel sounder. A dwell time of 1 ms was applied to the VNA measurements to ensure proper settling while taking measurements. The number of frequency points, N_{VNA}, taken by the VNA was computed from the spatial resolution of the channel sounder.

We compute the bandwidth-limited, measured impulse response, $\mathrm{h}_{\mathrm{VNA}}(t)$, and power delay profile, $\mathrm{PDP}_{\mathrm{VNA}}(t)$, of the channel by taking an average of the calibrated transmission parameters, S_{12} and S_{21}, assuming the channel is reciprocal:

$$\mathrm{h}_{\mathrm{VNA}}(t) = \mathit{IFFT}\left(\frac{\mathrm{S}_{12}(f)+\mathrm{S}_{21}(f)}{2}\right). \qquad (1)$$

$$\mathrm{PDP}_{\mathrm{VNA}}(t) = |\mathrm{h}_{\mathrm{VNA}}(t)|^2. \qquad (2)$$

The VNA path gain, G_{VNA}, can be computed by averaging over the frequency-domain data. Note that the path gain in this work does not include antenna gains since the channel included only coaxial cables and attenuators. For a VNA measurement, the path gain may be computed from the calibrated frequency-domain *S*-parameters as:

$$G_{\mathrm{VNA}} = \left(\frac{1}{\mathrm{N}_{\mathrm{VNA}}}\sum_{n=1}^{\mathrm{N}_{\mathrm{VNA}}}\left|\frac{\mathrm{S}_{12}(f)+\mathrm{S}_{21}(f)}{2}\right|^2\right)^{1/2}. \qquad (3)$$

Jargon, Jeffrey; Quimby, Jeanne; Remley, Kate; Koepke, Amanda; Williams, Dylan. "Verifying the Performance of a Correlation-Based Channel Sounder in the 3.5 GHz Band with a Calibrated Vector Network Analyzer." Paper presented at Asia-Pacific Microwave Conference (APMC) 2019, Singapore, SG. December 10, 2019 - December 13, 2019.

IV. Measurement Comparison

Despite the VNA and the correlation-based channel sounder sharing the same conducted channel, imperfections in the switches and differences in cable lengths used by the VNA and channel sounder meant that the two instruments were connected to slightly different channels. To overcome this, the reference planes of the VNA measurement were shifted to those of the channel sounder in post-processing. Having characterized the *S*-parameters of the switches, de-embedding and embedding procedures were used to shift the reference planes of the VNA to the channel sounder's reference planes, thus enabling a direct comparison of the two instruments [9].

Table 1 lists the correlation-based channel sounder and the VNA path gains for both the direct and multipath channels. A variable attenuator was switched among three values (approximately 18 dB, 28 dB, and 38 dB) for each channel. The standard uncertainties accompanying the VNA measurements include components due to both systematic and random effects, while the standard uncertainties for the correlation-based channel sounder only include components due to random effects. The nominal values and standard uncertainties were calculated from five repeated sets of five measurements taken in rapid succession each day for a duration of five days. For comparison purposes, the differences are also tabulated in column 3, as are the root-sum-of-squares (RSS) of the uncertainties. The data show the differences are within 0.25 dB for all cases, and the path gains measured by the VNA are always slightly lower. The uncertainties for the differences are always less than or equal to the actual differences, which is likely because the standard uncertainties for the correlation-based channel sounder do not include components due to systematic effects, such as impedance mismatches between components, and frequency-dependent distortions in the transmitter and receiver.

Figure 2 illustrates the PDPs of the correlation-based channel sounder and VNA for multipath channel 1. The peaks at 49 ns and 104 ns are aligned to within 2 dB. Additionally, the noise floor of the channel sounder is considerably higher than that of the VNA.

V. Conclusions

We verified the performance of a correlation-based channel sounder in the 3.5 GHz band with a calibrated vector network analyzer by comparing measurements in two conducted propagation channels at the same reference planes. We estimated the components of uncertainty due to systematic effects of the VNA and random effects of both systems and found that the channel sounder and VNA measurements agreed to within 0.25 dB for values of path gain, and the peaks were aligned to within 2 dB for values of PDP.

Table 1. Comparisons of path gain between the correlation-based channel sounder and the VNA.

Direct Channel		Path Gain ± Std. Unc. (dB)	Difference ± Unc. (dB)
1	VNA	-53.52 ± 0.13	0.14 ± 0.14
	Channel Sounder	-53.38 ± 0.05	
2	VNA	-63.38 ± 0.03	0.25 ± 0.08
	Channel Sounder	-63.13 ± 0.07	
3	VNA	-73.43 ± 0.09	0.16 ± 0.10
	Channel Sounder	-73.27 ± 0.05	
Multipath Channel		**Path Gain ± Std. Unc. (dB)**	**Difference ± Unc. (dB)**
1	VNA	-60.63 ± 0.12	0.24 ± 0.14
	Channel Sounder	-60.39 ± 0.07	
2	VNA	-70.56 ± 0.07	0.14 ± 0.11
	Channel Sounder	-70.42 ± 0.09	
3	VNA	-80.58 ± 0.11	0.25 ± 0.15
	Channel Sounder	-80.33 ± 0.10	

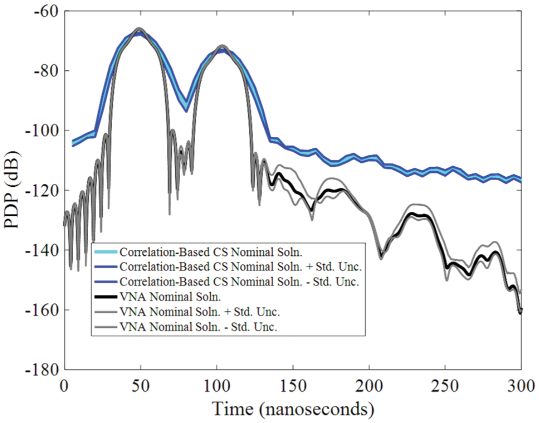


Fig. 2. Comparison of power delay profiles of the correlation-based channel sounder and VNA for multipath channel 1.

References

[1] S. Salous, "Radio Propagation Measurement and Channel Modeling," John Wiley & Sons Ltd., 2013.

[2] "CBRS Operational and Functional Requirements," Wireless Innovation Forum, Document WINNF-TS-0112, Version V1.7.0, May 2019.

[3] J. T. Quimby et al, "Channel Sounder Measurement Verification: Conducted Tests," soon to be published as NIST Technical Note.

[4] A. Molisch, "Wireless Communications," J. Wiley & Sons Ltd., 2011.

[5] D. F. Williams and B. Jamroz, NIST Microwave Uncertainty Framework, Beta Version, www.nist.gov/services-resources/software/wafer-calibration-software, 2017.

[6] J. T. Quimby et al, "NIST Channel Sounder Overview and Channel Measurements in Manufacturing Facilities," Technical Note (NIST TN) 1979 (2017).

[7] A. Goldsmith, "Wireless Communications," Cambridge University Press, 2005.

[8] J. A. Jargon, D. F. Williams, and P. D. Hale, "Developing Models for Type-N Coaxial VNA Calibration Kits within the NIST Microwave Uncertainty Framework," *87th ARFTG Conference*, May 2016.

[9] D. M. Pozar, "Microwave Engineering," John Wiley & Sons, Inc., 2005.

Jargon, Jeffrey; Quimby, Jeanne; Remley, Kate; Koepke, Amanda; Williams, Dylan. "Verifying the Performance of a Correlation-Based Channel Sounder in the 3.5 GHz Band with a Calibrated Vector Network Analyzer." Paper presented at Asia-Pacific Microwave Conference (APMC) 2019, Singapore, SG. December 10, 2019 - December 13, 2019.

Green Button Data-Access Model for Smart Cities

Lessons Learned on Security, Transfer, Authorization, and Standards-Compliance in Sharing Energy & Water Usage Data

Cuong Nguyen
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
cuong.nguyen@nist.gov

Jeremy J. Roberts
Green Button Alliance, Inc.
Raleigh, North Carolina, USA
jroberts@greenbuttonalliance.org

ABSTRACT

This paper provides a case study and lessons learned through the roll-out of the U.S. National Institute of Standards and Technology and U.S. Department of Energy Green Button electricity, natural gas, and water data-access initiative: to make readily available energy and water consumption data for consumers and third-party companies assisting mutual customers of utilities while protecting the security and privacy of the data. Energy and water usage data are important for smart cities in addition to individual consumers. Smart-city solutions rely heavily on the availability of such data to provide situational awareness as well as to inform control, actuation, and decision-making processes. However, the data need to be protected both for security and integrity. This paper presents a case study using the Green Button standard and the steps taken to ensure data security and privacy while enabling access to those consumption data by the consumer and third parties. Data security and privacy were achieved through use of the Green Button standard and subsequent implementation by the Green Button Alliance of a compliance-testing program. Considerations and solutions were needed for data in transit, data at rest, and the authorization mechanisms for allowing unregulated third-party companies to interface directly to utilities on behalf of the consumer while ensuring the consumer maintains complete control of what is to be shared and the ability to revoke that sharing at any time. The lessons learned from this approach could be applicable to other smart-city data.

CCS CONCEPTS

• Information systems → Data management systems → Data structures → Data access methods; • Information systems → World Wide Web → Web services → RESTful web services; • Security and privacy → Database and storage security → Data anonymization and sanitization

KEYWORDS

Green Button, Energy and Water Usage Data, Data Sharing, Data Security, Data Standard, Compliance Testing

ACM Reference format:

Cuong Nguyen and Jeremy J. Roberts. 2019. Green Button Data-Access Model for Smart Cities: Lessons Learned on Security, Transfer, Authorization, and Standards-Compliance in Sharing Energy & Water Usage Data. In *Proceedings of The 2nd ACM/EIGSCC Symposium On Smart Cities and Communities, 10 September 2019 (SCC 2019). ACM, Portland, OR, USA, 6 pages.* https://doi.org/10.1145/3357492.3358624

1 INTRODUCTION

The Green Button initiative [1] was led and implemented by the U.S. National Institute of Standards and Technology (NIST), the U.S. Department of Energy (DOE), and industry partners in response to a White House call-to-action [2] to provide consumers with machine-readable energy-usage information in a standard electronic format. The ultimate goal of the initiative was to build an ecosystem that enables utility customers to have easy and secure access to their energy-usage information in a consumer-friendly and machine-readable format for electricity, natural gas, and water usage, and to readily and securely share this data with partners identified by the consumer. NIST identified energy usage information as a priority in its coordination effort with industry on smart grid interoperability standards under the Energy Independence and Security Act of 2007 [3]. The Green Button initiative leveraged the development of a standard for energy usage information. NIST led this effort with the public-private partnership known as the Smart Grid Interoperability Panel (SGIP)—now a part of the Smart Electric Power Alliance (SEPA)—that completed the North American Energy Standards Board (NAESB) REQ18/WEQ19 Energy Usage Information and REQ21 Energy Services Provider Interface (ESPI)/Green Button technical standards [4]. The Green Button standard(s) provided an extensible markup language (XML) format for data exchange. Beyond the standard, NIST (David Wollman, Martin Burns, and John Teeter) led the development team that worked on open-source reference implementations, test tools, and technical artifacts to support the creation of a Green Button ecosystem, and supported initial utility Green Button implementations in California and elsewhere. The next major step for the effort was to create an organization that would maintain the development and evolution of the standard and manage-and-grow the ecosystem. The Green Button Alliance (GBA) was formed as a non-profit organization with support from NIST, DOE, and industry partners.

There are two main methods for utility customer to obtain their Green Button data: Download My Data (DMD) and Connect My Data (CMD). DMD provides a mechanism for the customer to view and monitor their energy-usage information directly from a utility website or portal. CMD allows the customer to share their usage information with a third-party service provider, such as an

Nguyen, Cuong; Roberts, Jeremy. "Green Button Data-Access Model for Smart Cities." Paper presented at The 2nd ACM/EIGSCC Symposium On Smart Cities and Communities, Portland, OR, US. September 10, 2019 - September 12, 2019.

application developer. One of the main focuses of the development was to ensure data security and privacy, to help address concerns that energy-usage data can be disaggregated to show the activities within a home; including whether the occupants are present [5]. Another focus of the effort was to create a testing program to assure that Green Button implementations conform to the standard, to improve interoperability within the Green Button ecosystem. The last major focus was to grow the Green Button ecosystem including to enhance the adoption and maintenance of the standard.

2 APPROACH TO PLATFORM STRUCTURE AND PLANNING

Significant planning and consideration went into determining the structure of energy-usage data. It was imperative to choose or create a format that would allow for future enhancements with backward compatibility. Electricity-usage data were the first data envisioned to be carried by Green Button, named after a government initiative for the sharing of Veteran Administration medical history information downloading known as the Blue Button [6]. Because the Green Button standards also included capability to handle other energy types, expansion of Green Button to include natural-gas data was quickly added, followed by water-usage data. Billing information and the concept of a split-and-parallel stream of data to carry personal information separate from the energy usage data were also implemented. Anticipating that there would be later additions—that were unknown at the time of inception—XML was deemed to be the most-flexible and readily available data-formatting language. It is one that could also be used with off-the-shelf tools and unaltered Web servers. The use of XML provided for a customer-friendly format, whereby the data are digital but can be read in a 'self-defining tag' system that allows people with an interest in the data to be able to see the data in a format that includes labels or tags without having to download or purchase a special parser. While XML is not visually friendly, it does allow for those who have an interest in understanding the underlying format to be able to view and interpret the data with an XML viewer or, with some difficulty, a text editor. Combined with XML Schema Definition (XSD) files—dictionary-like files that describe the format of the XML elements (tags), limits of values, and data types of values—most of the context of the data can be surmised.

The Atom Syndication Format, which acts as a wrapper to the Green Button -specific information, was selected to facilitate the ease of use in programming and conveyance of data using off-the-shelf tools that could already handle XML, since Atom itself is based on XML. Atom provided the ability to have data streamed without a need for defining a relational database structure or other non-flat file format but to still achieve the benefits of such non-flat relationships of data-to-other-data. XML transfers easily and can be parsed by most web servers without customization or need for add-on parsers. Atom parsers are readily available as well.

The U.S. Federal Government, along with some State and Canadian provincial governments, encouraged utilities and vendors to work together to make the Green Button a reality. The initial utility implementations of the standard were used to identify needed iterations in the standard. The advantage of this parallel activity was that the standard could be modified while it was being developed as things were discovered necessary or unnecessary in the implementations. The disadvantage was that there was no "gold standard" or "litmus test" for the implementations to mirror for interoperability. One lesson learned from this approach was that the latter disadvantage turned out to be greater than the advantage of parallel efforts due to the inability for third parties to create a single tool that would read and interpret data from multiple utilities without customization or tweaking of those data files. For most small companies, this proved too much of an effort to overcome with their limited resources.

Enhancements now (as then), come to the standard through an open workgroup known as the OpenADE Task Force (where ADE refers to automated data exchange) [7]. With industry and NIST leadership of OpenADE Task Force, both within the Utility Communications Architecture International Users Group (UCAIug) and its current home in the GBA, an effective forum was created to identify and track requested standards improvements. The OpenADE Task Force participation requires no membership, fees, commitments, or registrations by any company or individual. The results of these initial and subsequent enhancement efforts were given to the NAESB Energy Services Provider Interface (ESPI) Task Force (TF), a group of their Retail Electric Quadrant (REQ) and Retail Gas Quadrant (RGQ), which are focused on issues impacting the retail sale of energy to Retail Customers [8]. The ESPI TF was, and still is, used as the mechanism to standardize Green Button enhancements which are published and known officially as NAESB REQ.21 ESPI [9].

3 APPROACH TO THE PROTECTION OF PRIVACY/SECURITY

Security of customer energy-usage data and any personally identifiable information (PII) was a key and critical component of defining the solution for data in transit. While the Green Button standard scope of effort does not cover how data are to be stored at the utility nor at the third-party providing services for a mutual customer (data in situ), Green Button focuses on the security of these data in terms of their authorization and deliverance.

Green Button emphasizes five core tenets:

- multiple streams of data,
- adherence to modern web-transit standards,
- verification of party identity,
- the authority of the data custodian for customer verification, and
- the concept of customer consent and control.

3.1 MULTIPLE STREAMS OF DATA

As part of an approach to ensure that man-in-the-middle attacks do not breach entire sets of data, data in transit are separated into two streams: energy-usage information (EUI) and PII. In this way, any breach on one of the two data streams would not reveal the content of the other data stream. The information is obtained out of context of the other and usage data without context to whom it belongs, or personal information, like an address, without the associated usage values. Figure 1 below illustrates the separate streams of data. Security of customer energy-usage data and any personally identifiable information was a key and critical component of defining the solution for data in transit.

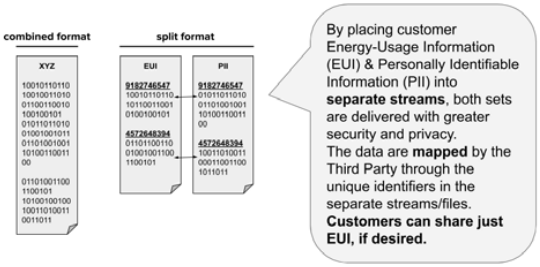


Figure 1: Split Data Format

3.2 ADHERENCE TO MODERN WEB-TRANSIT STANDARDS

In addition to the separation of data streams, security transit is ensured by the reference of NIST Federal Information Processing Standards (FIPS) 140-2 L1 cybersecurity standards [10] and use of the latest cipher suites on both ends—sending and receiving—and through the use of TLS 1.2 (or greater) and Certificate Authority - issued web certificates for the transfer of data. Figure 2 below illustrates the protection scheme for the data transfer.

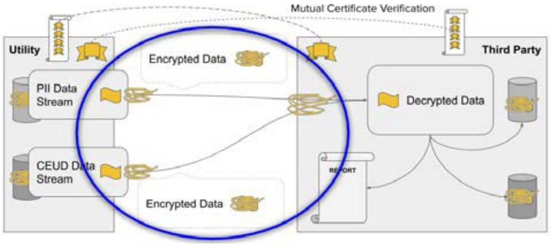


Figure 2: Encrypted Transfer

3.3 DOWNLOAD VS. CONNECT

As mentioned in an earlier section, the Green Button ecosystem is comprised of two different methods for obtaining data: (1) for a customer to download their data after defining the parameters of the scope of that data set and (2) for a customer to connect their data directly from the utility to a third- party provider after defining the scope of the data set. In the former method, DMD, there is no Alliance- or standards-defined way for obtaining these data; only for the format of these data in terms of the file structure. In the latter method, CMD, the Green Button defines the handshaking and exchange of these data in addition to the file (or stream) structure. While the workflows are different, the end goal is the same: that the customer is provided their data—either before analysis in the case of DMD or after analysis in the case of CMD.

Since DMD is nearly a subset of CMD, verification of party identity and the use of authorization are only necessary for CMD. There are inherent benefits of both methods—with DMD being easier from the standpoint of security and conveyance of data and CMD being more robust for continued access to data—but overall, CMD provides for an easier user experience at the expense of a greater development effort on the side of both the utility and the third-party provider for CMD deployment. Further, the workflows have different starting and ending paths, which must be considered in the development of a Green Button platform.

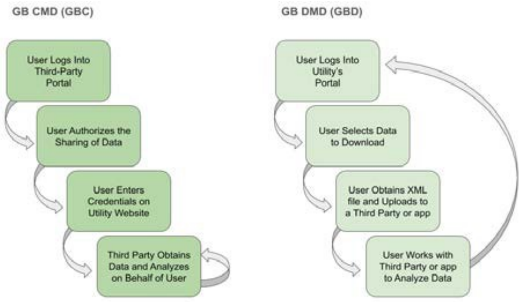

GB CMD (GBC)
User Logs Into Third-Party Portal
User Authorizes the Sharing of Data
User Enters Credentials on Utility Website
Third Party Obtains Data and Analyzes on Behalf of User
GB DMD (GBD)
User Logs Into Utility's Portal
User Selects Data to Download
User Obtains XML file and Uploads to a Third Party or app
User Works with Third Party or app to Analyze Data

Figure 3: Workflow of CMD and DMD

As shown in Figure 3, the workflow for CMD begins with a customer starting at a third-party provider's website where they would select their utility from a menu of utilities for which the third-party provider has a relationship; what is known as the third-party provider being already on-boarded with that utility (having met the technical and legal requirements set forth by the utility and/or jurisdictional authorities). The third-party provider would then direct the customer to the utility website with an application programming interface (API) call: essentially a web link and associated parameters. The API call would include the desired scope (the type of data, historical amount, interval, etc.). Subsequently, the utility would present the customer with verification screens for authentication (proof of the customer identity) and authorization (agreement to share the data scoped by the parameters in the API call). When complete (successful), the customer would then be sent back to the third-party provider's website using the provided return web link to complete the relationship with that provider.

Everything else is handled behind the scenes: the sharing of unique "tokens" for the establishment of the relationship in addition to the subsequent and ongoing data exchanges (more on that later). The customer would then utilize the services of the third-party provider for understanding their data.

The workflow for DMD begins by the customer/user logging into a utility's customer portal where they would select the data they wish to download, would obtain that data set as an XML file (or multiple XML files), and would then leave the utility's customer portal to go to a third-party providers portal or application where they would upload these data sets for their desired analysis and interpretation of their data.

3.4 VERIFICATION OF PARTY IDENTITY

Because the workflow for CMD includes the handshaking between utilities and third-party providers, verifying the identity of the other party can be an important addition to the security toolbox. Both utilities and third-parties can verify and prove the identity of their interfacing party by keeping repositories of each other's public certificates to ensure that the certificate chain is intact and that the entity interacting with them is that which is expected. This repository or databasing of certificates (or the databasing of digital signatures/thumbprints of a certificate) is an out-of-band exercise and thus by being out of band, can provide additional security at the expense of a manual process when those certificates change and need to again be shared with the interfacing party. The party-identity verification, as shown in Figure 4, would take place before any transfer of data between the utility and third-party to help in ensuring that the data in transit are in fact between those two entities.

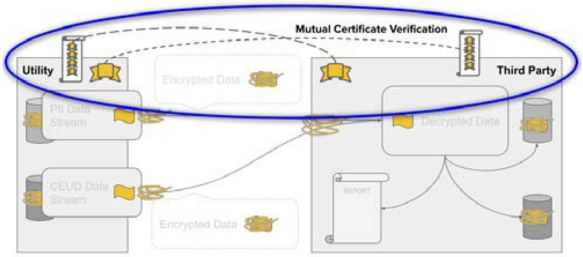


Figure 4: Mutual Certificate Verification

While data-at-rest is out of scope for Green Button, GBA and NIST have worked closely with the U.S. Department of Energy's DataGuard Energy Data Privacy Program [11] as a partner in promoting their *à la carte* menu of options available to utilities and third parties—as well as to commissions and jurisdictions—for ensuring the safeguarding of data in situ. GBA also recommends the separated EUI and PII in transit be kept separated in situ to keep a potential security breach of EUI confined to EUI without ownership (PII) and a potential security breach of PII confined to PII without context (EUI). Further, GBA recommends encryption- at-rest at both the utility and the third party; with independent, non-shared keys. That is, third-party providers would have no keys to the utilities' databases and utilities would have no keys to the third-party providers' databases.

3.5 THE AUTHORITY OF THE DATA CUSTODIAN FOR CUSTOMER VERIFICATION

In consideration of authorization for access, it was determined that the primary relationship for data sharing is between the utility (the Data Custodian) and the customer; more so than between the customer and the third-party, because the utility already has pre-established relationships with customers that include knowledge of their verified physical domicile and contact information. Therefore, it was decided for Green Button that the utility would act as an identity authority for data-sharing authorization by the customer; certainly, for DMD (as it is the utility's portal that the customer navigates) but as well for CMD representing the sharing of data with a third-party provider.

It has become commonplace for Google, Facebook, LinkedIn, and other online, identity-based companies to act as identity authorities for the ease and security of creating online accounts with disparate companies across the web; that is, to use, for example, LinkedIn credentials to create an account, log into that account, and share information with a website unrelated to LinkedIn—a website of a third-party provider—rather than creating a brand-new and separate account with that provider. The method for doing that is the Open Authorization (OAuth) [12] where authorization is granted by the customer for a defined scope of access and that authorization/scope combination is represented by a unique "token" used for subsequent reference of the authorization and its scope. OAuth 2.0 (the latest OAuth version at the time of this paper) is an ongoing effort of the Internet Society's Internet Engineering Task Force (IETF) OAuth Working Group [13].

3.6 THE CONCEPT OF CUSTOMER CONSENT AND CONTROL

Similar to the online identity-based social-media companies, utilities act as identity authorities for the sharing of electricity, natural gas, and water information with third-party companies that will be serving their mutual customers. Therefore, a customer can request the sharing of information with the third-party by authorizing that third-party access data on their behalf; data that are restricted by a given scope that can include the type of data (electricity, natural gas, and/or water), the interval of reading (e.g., monthly, daily, hourly), the level of personal information (e.g., service address, meter number, only usage data), and the duration of the authorization (e.g., one year, two years) if customer-initiated or utility-initiated revocation does not take place sooner. The result of that OAuth authorization is the unique token that is shared between the utility and the third-party provider of services—and no utility's customer-login information is shared. Thus, the sharing of data is always subject to customer consent and the customer always retains the right to determine the level of that sharing with the ability to revoke that relationship at any time. Revocation of third-party-provider access by the mutual customer occurs at the utility via OAuth.

Use of OAuth allows for yet-another-set of off-the-shelf tools in the Green Button ecosystem; making deployment on implementations easier and the user experience across the web more consistent [14]. However, today, not all methods of creating OAuth tokens are secure: The Green Button standard does not allow the use of the OAuth 2.0 "Implicit" method of token creation; a method which has enabled several known cybersecurity breaches. Further, no use of the OAuth "Resource Owner" method is allowed for Green Button; a method which allows the use of User ID and Password for authorization of tokens—a potential insecurity.

3.7 DESTRUCTION AND BACKUP OF DATA

Although not within the scope of Green Button, it is also important to consider the full data lifecycle including destruction and backup of data: Any customer- initiated destruction of third-party provider collected data would need to be ensured via local regulation or via third-party provider contract with the utility at the time the third-party provider was on- boarded as a viable party to transact with the utility's mutual customer. The availability/ability for a customer to download a backup of data—from either the utility or the third-party provider— should be instituted by local regulation and set forth to be made available in the standardized Green Button XML file format to allow comparison and/or porting of the data.

4 GREEN BUTTON TESTING AND CERTIFICATION PROGRAM

Testing and certification (T&C) program development is another critical component of the Green Button ecosystem. The standard provides the specifications and requirements for implementation. However, testing is needed to ensure that the product is implemented correctly in accordance with the standard's requirements. There are three types of T&C programs [15]. First-party certification is often known as "self-certification," where a manufacturer will attest that the product meets the requirements of the standard. Second-party certification is when a user tests and certifies the product; and in the case of the smart grid, it is mostly the utility that serves this role. Third-party certification is done through an independent authority that includes a certification body and associated test lab. The Green Button T&C was developed under the third-party structure with the UCAIug as the certification authority (administered by the GBA) and UL serving as both the certification body and the test lab. Most recently, GBA has taken the role of the certification authority and can also conduct testing, with the goal of the program being a cost- sensitive examination of implementations for compliance to the standard.

The T&C program-development effort was conducted in parallel with the standard development-and-enhancement to ensure that feedback from initial implementations could be incorporated into the standard. One effective approach was to develop the testing tools in an open-source environment that allowed the interested parties to contribute and use them. This has helped to speed up the testing of initial implementations. Also beneficial in driving the need for T&C development was information coming from hack-a-thons—gatherings of application programmers with the goal of determining interoperability between suppliers of data and readers of data.

4.1 LESSONS LEARNED

As mentioned in the discussion of the standardization effort, utilities (and vendors to utilities) were developing their Green Button implementations in parallel with the initial standardization effort. This resulted in DMD implementations that were producing customer usage files that were inconsistent with the format proposed in the standard XSD and inconsistent with the output of other utilities. In concert with the DOE and NIST, hack-a-thons were held to see how communities of interested parties would build the Green Button ecosystem.

The hack-a-thons had demonstrated that there needed to be more than standardization to ensure interoperability among implementations so that customization per utility was not necessary. While utilities were producing a file that would be used in a limited capacity geographically, third parties were expecting that their solutions could be rolled out to numerous customers in multiple geographies across North America and this was proven to be onerous without a way to ensure similar implementations by the utilities and the utilities' vendors.

It was upon the realization that there needed to be (a) a concerted and singular place for standardization efforts, (b) a way to verify and certify that implementations were compliant to the standard, and (c) a go-to place for finding reference implementations, technical support, educational materials, and collaboration, that the idea of a nonprofit organization to support the Green Button was conceived.

5 ECOSYSTEM DEVELOPMENT

Having a robust ecosystem is very important for any technologyto maintain development and enhance adoption of that technology. Through the efforts of the initial participants, and those that were interested in unified implementations, the GBA was launched in February 2015 to provide these services. The Green Button ecosystem was conceived to be a public-private initiative that would serve as an initial model for subsequent public-private initiatives for the collective benefit of American citizens and be a part of smart-city solutions. The Green Button initiative is seen as a way to take private, corporate and individual goals under the wing of the government to foster the efforts into a single focusand then to spin it out into its own initiative of self-preservation and growth. The creation of the nonprofit GBA was the culmination of those efforts and the handoff of government assistance; allowing the industry to grow in the hands of the corporate and individual participants in a concerted and unified manner. The GBA continues these initial missions through the support of grants and primarily, through the membership of interested parties.

To date, the Green Button ecosystem has grown to include the United States, Canada, and the Republic of Korea; and it is being considered by other countries as a model for their own energy-data-sharing programs. Figure 5 below shows the geographic extent of the Green Button ecosystem. The GBA and its partners continue to work with interested parties to grow the ecosystem and encourage further adoptions.

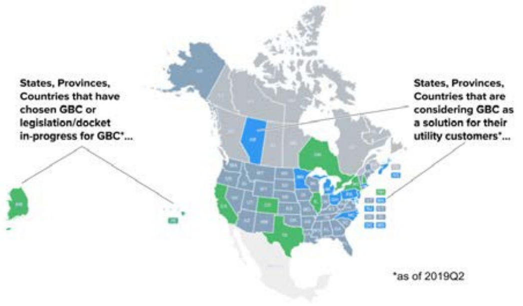

States, Provinces, Countries that have chosen GBC or legislation/docket in-progress for GBC*...
States, Provinces, Countries that are considering GBC as a solution for their utility customers*...
*as of 2019Q2

Figure 5: Ecosystem Map

6 CONCLUSION

The Green Button initiative for electricity, natural gas, and water data-access focused on freeing up consumption data. It provides a case study for the development and maintenance of a data-centric ecosystem. The first step of the initiative was a focus on the platform through the selection of XML as the data format. The next step was to standardize the requirements. This has helped with the initial implementation of products. A key aspect of the design was to ensure data security and privacy. The approach taken in Green Button is effective in ensuring security and privacy for the sharing of consumption data. The next important step was to develop a T&C program to assure that the implementation would be conformant to the standard. The last step was to form an ecosystem with a lead organization to further develop the technology and encourage further adoption. The lessons learned from this approach could be applicable to other smart-city data efforts as a model that addresses authorization, consent, and control in addition to cybersecurity when data access involves regulated and non-regulated entities' handling and sharing of data.

ACKNOWLEDGMENTS

Portions of this publication and research effort are made possible through the help and support of NIST via cooperative agreement 70NANB18H271. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose. The authors also acknowledge the significant contributions of NIST colleagues David Wollman and Martin Burns in their leadership of the Green Button initiative.

REFERENCES

[1] M.J. Burns, J.A. Teeter, and D.A. Wollman. Green Button: Building an Interoperable Ecosystem. Energybiz, 2014.

[2] National Science and Technology Council, A Policy Framework for the 21st Century Grid: Enabling Our Secure Energy Future, https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/nstc-smart-grid-june2011.pdf.

[3] Energy Independence and Security Act of 2007 [Public Law No: 110-140], Sec. 1305, http://www.gpo.gov/fdsys/pkg/PLAW-110publ140/pdf/PLAW-110publ140.pdf.

[4] North American Energy Standards Board (NAESB) ESPI/Green Button Standard, http://www.naesb.org/ESPI_Standards.asp.

[5] F.G. Mármol, C. Sorge, O. Ugus, and G.M. Pérez. Do not snoop my habits: preserving privacy in the smart grid. IEEE Communications Magazine, 2012, 50(5), 166-172.

[6] Health IT Blue Button. https://www.healthit.gov/topic/health-it-initiatives/blue-button.

[7] OpenADE. http://osgug.ucaiug.org/sgsystems/OpenADE/default.aspx.

[8] The North American Energy Standards Board (NAESB) Retail Electric Quadrant. https://www.naesb.org//naesb-req.htm.

[9] NAESB Energy Services Provider Interface Model Business Practices. https://www.naesb.org/ESPI_Standards.asp.

[10] Federal Information Processing Standard (140-2). https://csrc.nist.gov/publications/detail/fips/140/2/final.

[11] DataGuard Energy Data Privacy Program. https://www.smartgrid.gov/data_guard.html.

[12] Open Authorization Framework. https://oauth.net/.

[13] Internet Engineering Task Force. https://www.ietf.org/.

[14] M.J Burns. How The Green Button Initiative Secured Its APIs With OAuth. ProgrammableWeb. https://www.programmableweb.com/api-university/how-green-button-initiative-secured-its-apis-oauth.

[15] ANSI/NEMA SG-IPRM 1-2016, Smart Grid Interoperability Process Reference Manual. https://www.nema.org/Standards/Pages/Smart-Grid-Interoperability-Process-Reference-Manual.aspx.

A Machine Learning Approach to the Estimation of Near-Optimal Electrostatic Force in Micro Energy-Harvesters

Masoud Roudneshin[1], Kamran Sayrafian[2], Amir G. Aghdam[1]

[1]Department of Electrical Engineering,
Concordia University
Montreal, Canada

[2] Information Technology Laboratory,
National Institute of Standards & Technology
Gaithersburg, MD, USA

***Abstract*—Wearable medical sensors are one of the key components of remote health monitoring systems which allow patients to stay under continuous medical supervision away from the hospital environment. These sensors are typically powered by small batteries which allow the device to operate for a limited time. Any disruption in the battery power could lead to temporary loss of vital data. Kinetic-based micro-energy-harvesting is a technology that could prolong the battery lifetime or, equivalently, reduce the frequency of recharge or battery replacement. Focusing on a Coulomb-Force Parametric Generator (CFPG) micro harvesting architecture, several machine learning approaches are presented in this paper to optimally tune the electrostatic force parameter, and therefore, maximize the harvested power.**

***Index Terms*—energy harvesting, wearable sensors, microgenerator, CFPG**

I. Introduction

Wearable and implantable medical sensors are considered a key component of future telemedicine systems, allowing clinicians to have remote access to real-time patient's data [1], [2]. These devices typically operate by using small batteries; therefore, frequent recharge or battery replacement might be necessary to keep the device functioning properly. Prolonging the lifetime of these batteries and reducing their frequency of recharge could have a paramount impact on their everyday use. This is especially important for implanted devices, where battery replacement is not easily possible.

Energy harvesting refers to the process of scavenging energy from external sources (ambient environment such as solar power, wind and kinetic energy) [3]. For wearable devices, kinetic energy can be a reliable solution for power generation in medical sensors. For cases of nonstationary vibrations (for example, as a result of the human body motion), Coulomb-force parametric generator (CFPG) architecture has been proposed as a promising solution to extract power form human movements. [4], [5]. In this type of system, a proof mass can move between upper and lower bounds $\pm Z_l$ as shown in Fig. 1. The summation of the device motion to the inertial frame $\xi(t)$ and the relative motion of the proof mass with respect to the device $z(t)$ make the absolute motion of the proof mass equal to $y(t) = \xi(t) + z(t)$.

To model the dynamics of the proof mass motion in CPFG, the following nonlinear differential equation was proposed in [6]:

$$m\ddot{y}(t) = -m\ddot{z}(t) - F \times \text{Relay}(z(t)) \tag{1}$$

where m is the proof mass, $\ddot{y}(t)$ is the acceleration of the frame of CPFG with respect to the inertial frame, z is the relative acceleration of the proof mass with respect to the frame of CFPG, and F is the electrostatic holding force which acts against the motion of the proof mass. The generated mechanical power of the system is equal to the product of the electrostatic holding force and the relative velocity of the proof mass with respect to frame and is calculated as:

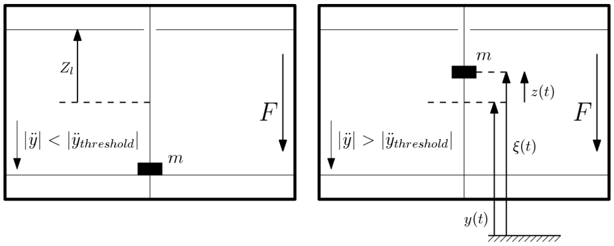

Z_l
F
$|\ddot{y}| < |\ddot{y}_{threshold}|$
m
m
$z(t)$
F
$|\ddot{y}| > |\ddot{y}_{threshold}|$
$\xi(t)$
$y(t)$

Fig. 1. Generic Model of a CPFG: (a) proof mass attached to one end, and (b) proof mass in flight

$$P(t) = F \times \dot{z}(t) \tag{2}$$

Fig. 1 displays the process of energy generation in CPFG. The proof mass is initially located at either upper or lower plates. The mass does not move until the external acceleration exceeds a certain limit. To harvest power, the external acceleration must be strong enough to create a full displacement of the mass from one plate to the other. If the mass cannot complete a full flight to the other plate, all generated power is consumed in the electrical field of CFPG by the electrostatic force F.

A further investigation of the power equation of CFPG coupled with its dynamics reveals how adaptation of F affects the output power of the generator [7]. The power is a function of both relative velocity $\dot{z}(t)$ and force F. Meanwhile, $\dot{z}(t)$, as evident from equation (1), is a function of both external acceleration $\ddot{y}(t)$ and force F. Therefore, an optimal strategy that ensures proper adaptation of the electrostatic force F leads to the maximization of the average output power exists. As such, the following optimization problem is considered:

$$\underset{F_i}{\text{argmax}} \left[\frac{1}{\Delta} \times \sum_{t=t_i}^{\Delta+t_i} P(t) \right], \tag{3}$$

with the constraints given by equations (1) and (2). In other words, it is desired to maximize the average harvested power

Sayrafian, Kamran; Roudneshin, Masoud; Aghdam, Amir. "Output Power Maximization of a Microgenerator Using Machine Learning Approach." Paper presented at The 7th Annual IEEE International Conference on Wireless for Space and Extreme Environments, Ottawa, CA. October 16, 2019 - October 18, 2019.

during the time interval $[t_0 + (i-1)\Delta, t_0 + i\Delta]$ by selecting the optimal value of the electrostatic force F_i.

The authors in [6] demonstrate that the output power of a CFPG micro-harvester can be maximized by proper adjustment or adaptation of the electrostatic force F. A methodology for optimizing F by observing the input acceleration in the previous time interval is also proposed in [6]. The average output power for different values of the holding force F and various locations of the wearable sensor is evaluated in [7].

In this paper, we propose a novel method for the estimation of the suboptimal value of the electorstatic force in a micro energy-harvester, according to the current absolute acceleration of the CFPG frame. We use the frequency spectrum of the human body acceleration data in our analysis. Eight different machine learning classification schemes are then used and their performances are compared in terms of accuracy in estimating the suboptimal value of the holding force in the next time step for power maximization. To the best of the authors' knowledge, this is the first time the frequency spectrum of the human body acceleration is used for power maximization in CFPG microgenerators.

The rest of the paper is organized as follows. In Section II, we illustrate our proposed method, and the procedure for generating artificial data is explained. Then in Section III, we discuss various methods of classification of labeled data. In Section IV, results for eight classification structures is obtained and their accuracy is discussed. Finally, conclusions are drawn in Section V.

II. Problem Definition

A. Acceleration in Human Body

The authors in [8] demonstrate that during normal daily activities, bulk of the frequency content of the human motion acceleration in the upper extremity is within the range 0.8-5Hz. Also, it is shown in [9] that 99% of the acceleration power spectral density, when walking barefoot, is concentrated below 15Hz. Based on these findings, we make the following assumption.

Assumption 1 *The acceleration signal for time intervals of length Δ can be approximated by the following cosine series:*

$$\ddot{y}(t) \approx \sum_{n=0}^{20} A_n(i)\cos(2\pi f_n t) = \sum_{n=0}^{20} A_n(i)\cos(2\pi n t), \quad t \in [t_0 + (i-1)\Delta, t_0 + i\Delta], \tag{4}$$

where $A_n(i)$ is the amplitude of the frequency component corresponding to f_n in the ith time interval. Having the amplitudes of the frequency components in every time interval, the problem reduces to identifying the mapping ϕ such that:

$$\phi : [A_0(i), ..., A_{20}(i)] \rightarrow \tilde{F}_\Delta(i), \tag{5}$$

The ultimate goal is to maximize the harvested mechanical power in equation (2) for time interval i by finding a pseudo-optimal holding force $\tilde{F}_\Delta(i)$. Given the time dependency of the parameters involved, we have divided this problem into two steps:

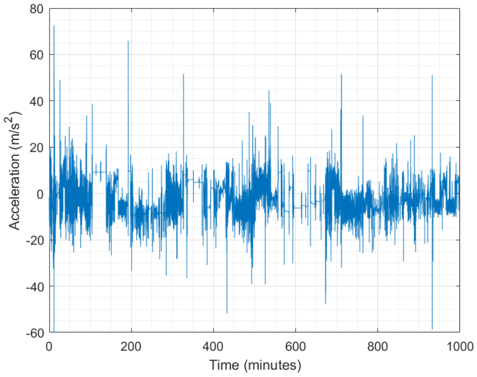


Fig. 2. A 1000-min acceleration sample of human arm motion

1) **Estimation of the pseudo-optimal holding force during time interval i:**
Assuming that we know the $A_n(i)$ coefficients for the acceleration signal in the time interval $[t_0+(i-1)\Delta, t_0+i\Delta]$, what is the pseudo-optimal holding force $\tilde{F}_\Delta(i)$ that maximizes average harvested power for the same time interval i.e. $[t_0 + (i-1)\Delta, t_0 + i\Delta]$?
2) **Estimation of the optimal interval size to maximize the average harvested power for the next time interval i.e. $[t_0 + i\Delta, t_0 + (i+1)\Delta]$:**
In practice, the information about $\tilde{F}_\Delta(i)$ can not be available for the current time interval $[t_0+(i-1)\Delta, t_0+i\Delta]$; and, at best, $\tilde{F}_\Delta(i)$ can be applied to the next time interval i.e. $[t_0+i\Delta, t_0+(i+1)\Delta]$. Therefore, depending on the temporal correlation of the estimated pseudo-optimal holding force, the achieved harvested power will be less than the value obtained in step 1. This step requires further study on optimizing the interval size Δ.

The focus of this paper is on solving the first step through machine learning algorithms. In the next subsection, we will present several methodologies for data classification in order to find a suitable mapping from the frequency spectrum of the acceleration to the electrostatic force F.

B. Acceleration Data Processing

The acceleration data of the human arm motion, as obtained in [10], is used in our analysis. Fig. 2 demonstrates a 1000-min data obtained by attaching an accelerometer to a male subject arm. The data is interpolated with 1ms sampling steps. Then, it is divided into intervals of length equal to one second for approximation with the cosine functions as stated in equation (4). To reduce the size of the action space, the values of the holding force F are divided into 1mN steps. Then, similar to the process outlined in [6] and [7], we choose a value for $\tilde{F}_\Delta(i)$ from the set $\{2, 3, ..., 10\}$.

To label each of the 1-sec intervals with their corresponding holding force which maximizes the harvested power in that

Sayrafian, Kamran; Roudneshin, Masoud; Aghdam, Amir. "Output Power Maximization of a Microgenerator Using Machine Learning Approach." Paper presented at The 7th Annual IEEE International Conference on Wireless for Space and Extreme Environments, Ottawa, CA. October 16, 2019 - October 18, 2019.

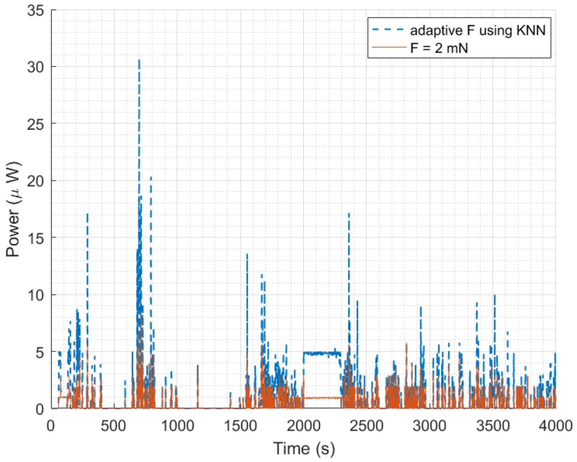


Fig. 3. Comparison of the harvested power for a 4000-sec data for the adaptive holding force with KNN algorithm and a constant holding force

time interval, an algorithm based on the work in [6] is implemented for each of the holding forces and the pseudo-optimal force is obtained accordingly. A set of 1950 labeled data for each of the nine classes (making a total of 17550 vectors) is obtained subsequently.

III. Data Classification

As discussed earlier, the estimation of the electrostatic holding force which maximizes the output power based on the spectral content of the acceleration signal as described in equation (4) can be studied in the context of a classification problem. In what follows, eight different classification schemes are briefly described.

- **Decision Tree Classifier [11]**
A decision tree is a flowchart-like tree structure. An internal node represents a feature (or attribute), each branch represents a decision rule, and each leaf node represents an outcome. The best attribute is selected using an appropriate attribute selection measure (ASM) such as information gain or Gini index. Then, the selected attribute is used as a decision node, the dataset is broken into smaller subsets, and these steps are repeated recursively until a matching condition is satisfied.

- **Random Forest Classifier [12]**
Random forest consists of a large number of individual decision trees that operate as an ensemble. A set of decision trees are created from a randomly selected subset of the training set. Then, the classifier aggregates the votes from different decision trees to decide the final class of the test object. Therefore, a classification is made based on the majority of votes received from each of the decision trees. It is to be noted that a single decision tree may be prone to noise, but aggregating many decision trees reduces the effect of noise, leading to more accurate results.
- **K-Neighbors Classifier [11]**
In this method, an object is classified by a majority vote of its neighbors, with the object being assigned to the most common class among its k nearest neighbors.
- **Support Vector Machine (SVM) [13]**
This is a discriminative classifier which is formally defined by a separating hyperplane. Given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new entries. The hyperplane can be a linear, polynomial or exponential function of the weighted sum of inputs, where the weighting coefficients are optimized in the training process.
- **Multi Layer Perceptron (MLP) [14]**
In this approach, each computational block consists of a weighting matrix that is multiplied by the inputs. The outputs are then passed to a function called *activation function*. These blocks are often repeated several times, and the output of each block is used as the input of the next one. Finally, the output, which is in the form of a vector, classifies the input in the form of an output vector. The elements of the weighting matrices are updated to minimize the error between the estimated vector and a label vector which is associated with the correct class.
- **Stochastic Gradient Descent (SGD) [15]**
This is a linear classifier just like a linear hyperplane in SVM. The only difference is that the optimization technique applied to find the weighting parameters of the hyperplane is stochastic gradient descent. To update the weights, the gradient of loss function is needed. For the computation of the gradient of the cost function, the sum of the cost of each sample is needed, which could be inefficient for large training datasets. On the other hand, when applying SGD, the cost gradient of only one sample is used at each iteration (instead of the sum of the cost gradients of all training datasets) which can significantly reduce the computational complexity of the algorithm.
- **Passive Aggressive Classifier [16]**
In this method, a linear function of the input multiplied by weighting elements is passed to an activation function (often a sign function for binary classifications). Then, a Hinge loss function is used to measure the error between the estimated and true values of labels. The update rule for weighting elements works in such a way that the algorithm is passive when a correct classification occurs (no weight change). For false classification cases, on the other hand, the algorithm becomes aggressive and updates the weights so as to minimize the loss for similar inputs that may occur in other instances.

- **Ridge Classifier [17]**
This method often uses a linear function of the input vector and weighting parameters. The function used for updating the weights is the squared error of the estimated

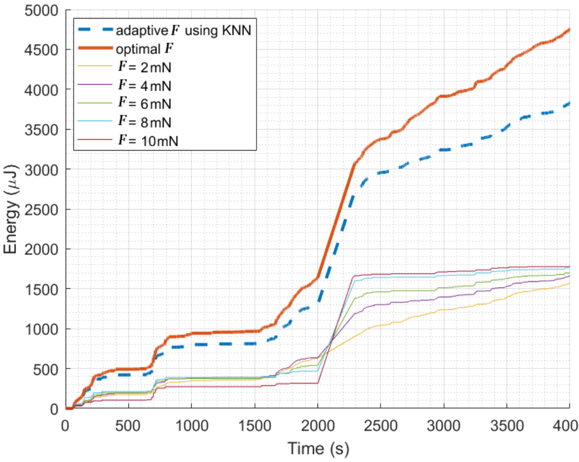


Fig. 4. Comparison of the harvested energy for a 4000-sec time sample for the ideal case with optimal value of F, adaptive holding force with KNN algorithm, and a constant holding force

outputs plus a regularization term which is a function of the weights. The regularization technique tends to reduce overfit in estimation.

IV. Simulation Results

The vectors containing the amplitudes of the cosine approximation for the nine classes of pseudo-optimal force $\tilde{F}_{\Delta}(i)$ are first divided by their maximum magnitude to normalize the vector elements between -1 and 1. Then, 90% of these 17550 vectors are randomly selected from each class in order to be used as the training data set, and the remaining 10% are used as the test data. All of the eight classification techniques described in Section III are compared by simulations, which are performed on a computer with an Intel R processor (Core i5) running at 2.5GHz using 16GB of RAM with Windows 10[1].

For training the classification algorithms and deploying machine learning models, we use the Google TensorFlow[1]platform. The results are given in Table I, indicating the running times for each input vector and accuracy of the learning methods in estimating the pseudo-optimal F. Fig. 3 shows the improvement in the generated power when the electrostatic force is changed using our approach, compared to the case of applying a constant force F=2mN. The amount of gain depends on many factors such as the dimension of the mass-spring-damper inside the CFPG micro-harvester. For these results, the size of the MSD and the distance between the two plates $(2Z_l)$ were chosen as 15×15×1.5mm^3 and 0.5mm, respectively.

TABLE I
Running time and accuracy of eight classification techniques for test data of the acceleration of human arm

	Running Time (s)	Accuracy (%)
Decision Tree	0.0009	45
Random Forest	0.3198	86
KNN	0.032	94
SVM	0.0866	73
MLP	0.0009	86
SGD	0.0010	43
Passive Aggressive	0.1149	50
Ridge	0.0001	43

Fig. 4 displays the gain in the harvested energy for a 4000-sec time interval. The harvested energy for different constant holding forces is also plotted. In addition, for this example, we have included the amount of harvested energy using the optimal value of the holding force as a reference. The curve corresponding to the optimal F serves as an upper bound for the amount of energy that can be generated. Although it is desirable for the adaptive holding force methodology to be as close as possible to the optimal curve, one should also take into consideration the computational complexity of the algorithms used. The higher this complexity is, the more energy it requires to perform, leading to less (or even no) gain in the harvested power. Therefore, a thorough analysis of the implementation complexity and the resulting power consumption of the added hardware is required to justify the addition of the optimization methodology to the micro-harvester circuitry. Our initial investigation shows that algorithms such as MLP can be implemented with reasonable runtime power consumption. Further details of our analysis will be provided in future publications.

V. Conclusion

Efficiency of eight machine learning classification techniques for adaptive estimation of the electrostatic force in a CFPG architecture according to the frequency spectrum of acceleration in human arm was investigated. It was observed that the net energy harvested using a suitable machine learning technique could lead to significant gain in the harvested power. The exact amount of the gain depends on many parameters such as the dimension of the MSD inside the harvester, adaptation time interval Δ, and the classification algorithm. The impact of these parameters will be studied in more details in future research.

Acknowledgement

This work was supported by the National Institute of Standards and Technology (NIST) [grant number 60NANB18D261].

References

[1] H. C. Koydemir and A. Ozcan, “Wearable and implantable sensors for biomedical applications,” *Annual Review of Analytical Chemistry*,vol. 11, pp. 127–146, 2018.

[1] Intel R Processor, Windows 10 and Tensorflow are products of Intel Corp., Microsoft and Google, respectively. These products have been used in this research to foster research and understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that this product is necessarily the best available for the purpose.

[2] Y. Khan, A. E. Ostfeld, C. M. Lochner, A. Pierre, and A. C. Arias,“Monitoring of vital signs with flexible and wearable medical devices,” *Advanced Materials*, vol. 28, no. 22, pp. 4373–4395, 2016.

[3] F. K. Shaikh and S. Zeadally, “Energy harvesting in wireless sensor networks: A comprehensive review,” *Renewable and Sustainable Energy Reviews*, vol. 55, pp. 1041–1054, 2016.

[4] P. D. Mitcheson, T. Sterken, C. He, M. Kiziroglou, E. Yeatman, and R. Puers, “Electrostatic microgenerators,” *Measurement and Control*, vol. 41, no. 4, pp. 114–119, 2008.

[5] T. Von Buren, P. D. Mitcheson, T. C. Green, E. M. Yeatman, A. S. Holmes, and G. Troster, “Optimization of inertial micropower generators for human walking motion,” *IEEE Sensors Journal*, vol. 6, no. 1, pp. 28–38, 2006.

[6] M. Dadfarnia, K. Sayrafian, P. Mitcheson, and J. S. Baras, “Maximizing output power of a CFPG micro energy-harvester for wearable medical sensors,” in *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies*, 2014, pp. 218–221.

[7] D. Budić, D. Šimunić, and K. Sayrafian, “Kinetic-based micro energy-harvesting for wearable sensors,” in *Proceedings of the 6th IEEE International Conference on Cognitive Infocommunications*, 2015, pp. 505–509.

[8] C. V. Bouten, K. T. Koekkoek, M. Verduin, R. Kodde, and J. D. Janssen, “A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity,” *IEEE Transactions on Biomedical Engineering*, vol. 44, no. 3, pp. 136–147, 1997.

[9] E. K. Antonsson and R. W. Mann, “The frequency content of gait,” *Journal of Biomechanics*, vol. 18, no. 1, pp. 39–47, 1985.

[10] N. Yarkony, K. Sayrafian, and A. Possolo, “Energy harvesting from the human leg motion,” in *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, 2014, pp. 88–92

[11] S. D. Jadhav and H. Channe, “Comparative study of K-NN, naive bayes and decision tree classification techniques,” *International Journal of Science and Research*, vol. 5, no. 1, pp. 1842–1845, 2016.

[12] L. Rokach, “Decision forest: Twenty years of research,” *Information Fusion*, vol. 27, pp. 111–125, 2016.

[13] D. Tomar and S. Agarwal, “A comparison on multi-class classification methods based on least squares twin support vector machine,” *Knowledge-Based Systems*, vol. 81, pp. 131–147, 2015.

[14] S. B. Wankhede, “Analytical study of neural network techniques: SOM, MLP and classifier-a survey,” IOSR *Journal of Computer Engineering*, ver. VII, vol. 16, no. 3, 2014.

[15] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT’2010*, 2010, pp. 177–186.

[16] J. Jorge and R. Paredes, “Passive-aggressive online learning with non-linear embeddings,” *Pattern Recognition*, vol. 79, pp. 162–171, 2018.

[17] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, “Extreme learning machine for regression and multiclass classification,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 513–529, 2011.

Using Statistical Methods and Co-Simulation to Evaluate ADS-Equipped Vehicle Trustworthiness

Khalid HALBA
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
khalid.halba@nist.gov

Edward GRIFFOR
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
edward.griffor@nist.gov

Patrick KAMONGI
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
patrick.kamongi@nist.gov

Thomas ROTH
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
thomas.roth@nist.gov

Abstract— With the increasing interest in studying Automated Driving System (ADS)-equipped vehicles through simulation, there is a growing need for comprehensive and agile middleware to provide novel Virtual Analysis (VA) functions of ADS-equipped vehicles towards enabling a reliable representation for pre-deployment test. The National Institute of Standards and Technology (NIST) Universal Cyber-physical systems Environment for Federation (UCEF) is such a VA environment. It provides Application Programming Interfaces (APIs) capable of ensuring synchronized interactions across multiple simulation platforms such as LabVIEW, OMNeT++, Ricardo IGNITE, and Internet of Things (IoT) platforms. UCEF can aid engineers and researchers in understanding the impact of different constraints associated with complex cyber-physical systems (CPS). In this work UCEF is used to produce a simulated Operational Domain Design (ODD) for ADS-equipped vehicles where control (drive cycle/speed pattern), sensing (obstacle detection, traffic signs and lights), and threats (unusual signals, hacked sources) are represented as UCEF federates to simulate a drive cycle and to feed it to vehicle dynamics simulators (e.g. OpenModelica or Ricardo IGNITE) through the Functional Mock-up Interface (FMI). In this way we can subject the vehicle to a wide range of scenarios, collect data on the resulting interactions, and analyze those interactions using metrics to understand trustworthiness impact. Trustworthiness is defined here as in the NIST Framework for Cyber-Physical Systems, and is comprised of system reliability, resiliency, safety, security, and privacy. The goal of this work is to provide an example of an experimental design strategy using Fractional Factorial Design for statistically assessing the most important safety metrics in ADS-equipped vehicles.

Keywords— ADS-equipped vehicles, cyber-physical systems, trustworthiness, co-simulation

I. Introduction and related work

The current state of modeling and simulation for vehicles includes proprietary tools, such as IGNITE [1], CANOE [2], and CANALYZER [3], and open-source tools such as OpenModelica [4]. [1] These tools include libraries to simulate vehicle dynamics such as steering, braking, energy conversion and transmission, and power management. Some of these tools allow for communication with external simulators through the FMI [5]. The ability to interface with external simulators widens the range of simulation scenarios that may include interactions with other components through control and sensing functions [6], [7]. In this work we use UCEF [8] to demonstrate that co-simulation can enable innovation in ADS-equipped vehicle research while allowing specialized simulation platforms to independently develop and improve traditional vehicle dynamics simulation.

In Section II we review the testbed functional components, their roles in our co-simulation, and their implementation. In Section III we demonstrate the potential of UCEF to implement the functional components in a co-simulation. In Section IV we propose a strategy for assessing safety metrics by defining an ODD and defining input and output parameters of interest using the Fractional Factorial Experiment Design [9]. The conclusion summarizes our effort and highlights future work.

II. ADS-equipped vehicles Testbed Functional Components and Implementation

In this section we describe the functional components of our ADS-equipped vehicle testbed and its implementation.

A. Functional Components

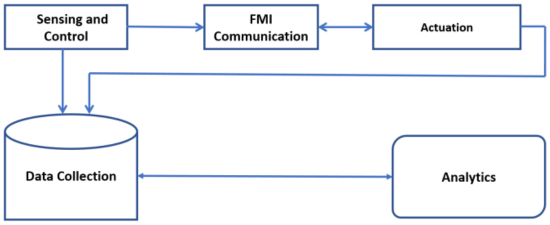


Fig. 1. ADS-Equipped Vehicles Testbed Functional Components

The ADS-equipped vehicle testbed is composed of the functional components described in Figure 1:

a) the Sensing and Control component responsible for receiving input data from the environment such as weather, road condition, traffic infrastructure, and events such as obstacle detection data. The control feature of this component reacts to the data received by

[1] Certain commercial products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Halba, Khalid; Griffor, Edward; Roth, Thomas; Kamongi, Patrick. "Using Statistical Methods and Co-Simulation to Evaluate ADS-Equipped Vehicle Trustworthiness." Paper presented at Electric Vehicles International Conference (EV) 2019, Bucharest, RO. October 03, 2019 - October 04, 2019.

the sensing feature and generates a corresponding drive cycle, i.e., an acceleration, and braking pattern. Decision support algorithms determine this behavior based on the given ODD;

b) the FMI Communication component plays a bridge role by transmitting the drive cycle produced from the sensing and control component to the actuation component;

c) the Actuation component receives the drive cycle produced by the sensing and control component then performs either acceleration or deceleration functions;

d) the Data Collection component collects data about the interactions between the sensing and control component and the actuation component for further processing and trustworthiness metrics assessment;

e) the Analytics component analyzes the data collected by the data collection component and runs statistical techniques and machine learning algorithms to assess trustworthiness metrics of the ADS-equipped vehicles testbed experiments.

B. *Implementation*

Now we describe the implemented features of the functional components. The implementation uses a NIST co-simulation platform called UCEF that is built using an Ubuntu virtual machine. UCEF relies on the concept of multiple federates that interact using a publish-subscribe message pattern to simulate different functions of a CPS. In the present work, UCEF is used to simulate ADS-equipped vehicle autonomy or decision-making functions focused on acceleration and deceleration.

a) the sensing and control functional component is implemented as two federates: i) a Sensing Federate that simulates a binary obstacle detection notification (vehicle or other obstacle) at a specific simulation time and sends that notification over the HLA bus to the Control Federate. ii) a Control Federate that processes the notification provided by the Sensing Federate and generates corresponding acceleration, and deceleration requests. The Control Federate also implements a User Datagram Protocol (UDP) server that sends drive cycle data to the actuation module over the communication component. Code for the Sensing and Control Federates is available in GitHub [10];

b) the communication component is responsible for communication between the UCEF federates and the Actuation component, built as Functional Mock-Up Unit (FMU). The FMU implements a UDP client that listens on the incoming drive cycle data from the UCEF UDP server implemented in the Control Federate and feeds that information to the Actuation ccomponent. We have built this FMU based on the Q.Tronic FMU Software Development Kit (SDK) and shared its source code in GitHub [11];

c) the Actuation component implementation can be realized using different simulators, including IGNITE, Modelica, and MATLAB that provide both libraries for running traditional vehicle functions and an FMI master algorithm that enables them to interact with other simulators. In this work we use Ricardo IGNITE, a vehicle simulation platform that can be installed on any Microsoft Windows computer, for modeling and simulation of electric and fuel-based vehicle models. IGNITE provides a built-in FMI Master algorithm, and models legacy vehicle functions based on the UCEF generated drive cycle;

d) the Data Collection component is implemented using a MySQL database UCEF federate. Each interaction, or message exchange, is time-stamped and represented in this database as a record that comprises data encoding, the sender, the receiver, the time of transmission, the time of reception, a description of the carried signal, and the payload of the message;

e) the Analytics component will be implemented in a future work as a federate to perform two functions: i) an assessment function using machine learning algorithms and Fractional Factorial Experiment Design scripts. In this work the Fractional Factorial statistical analysis is done using formulas implemented in a standalone excel data form ; ii) a Vert.X-backend [12] / Vue.js-frontend [13] microservice for data visualization was implemented and tested for visualizing post-experiment data collected by the Data Collection component.

III. Co-Simulation of the Functional Components

UCEF enables the co-simulation of a wide variety of CPS using the IEEE High Level Architecture (HLA), including systems at scale such as power grids [14]. Figure 2 shows a federation that will simulate message exchange for the sensing, control, communication, and analytics components using simulated J1939 CAN frames.

These federates were modeled in UCEF using the Web-based Generic Modeling Environment (WebGME), which includes JavaScript extensions to convert the models into Java code.

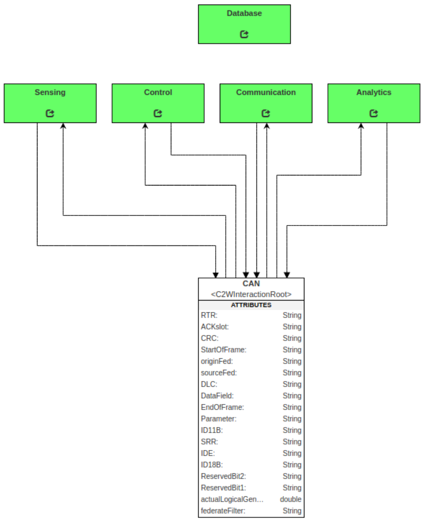


Fig. 2. ADS-equipped vehicles federation where the functional components exchange simulated CAN messages that are logged in a database

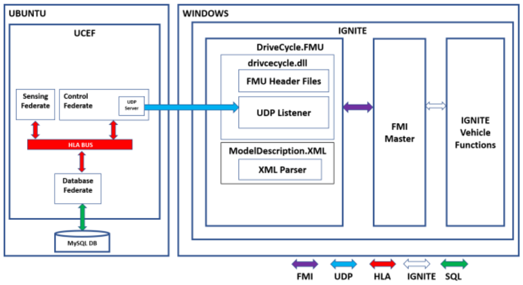


Fig. 3. UCEF implementation of ADS specific functions for an ADS-equipped vehicle testbed

WebGME was used to generate Java code for the model, and that code was implemented with the desired behavior for each federate. Java federates were modeled to represent the features of the functional components. The sensing federate sends obstacle detection data to the control federate over the HLA bus. The control federate adjusts the drive cycle based on the sensing federate input and generates a new drive cycle that will be fed to the actuation module through the FMU. The database federate collects and stores the simulated message exchange between the different federates.

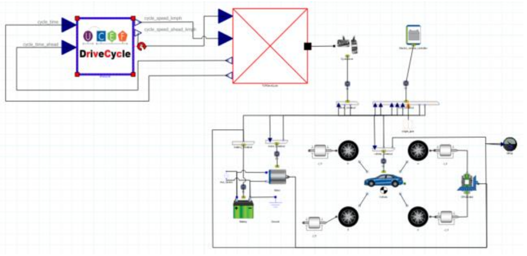


Fig. 4. Custom UCEF drive cycle loaded into IGNITE as an FMU that provides acceleration and braking signals to a simulated Electric SUV

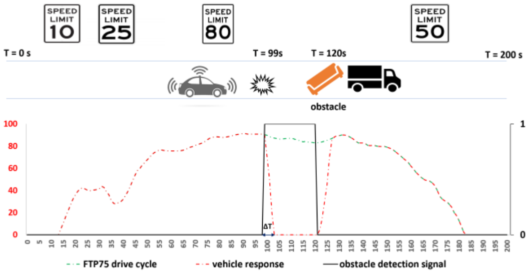


Fig. 5. ODD Overview with the simulated 200s of the FTP 75 drive cycle portion [150-350] fed to the Actuation Module. Stopping Time ΔT is also illustrated.

Other legacy vehicle simulators, such as IGNITE, have a mature implementation of traditional vehicle functions. As such it makes more sense for an ADS-equipped vehicle testbed to focus on extending the capabilities of these simulators by enabling interactions between them and environments like UCEF designed to simulate ADS or non-traditional vehicle functions. Figure 3 shows the integration of IGNITE with the sensing and control federation to produce a co-simulation of the of ADS-equipped vehicle testbed.

We have simulated an exchange of drive cycle data between UCEF and IGNITE. UCEF generates a portion of FTP75 [15] drive cycle data stream. The FMU implements a UDP listener. Figure 4 shows the integration of this external drive cycle data into the IGNITE simulation environment.

IV. Safety Study

In this section we present a scenario that demonstrates the potential of our ADS-equipped vehicles testbed in assessing trustworthiness of ADS-equipped vehicles. According to the National Highway Transportation Safety Administration NHTSA [16] and Waymo's Safety Report [17], an ODD refers to the conditions under which a self-driving system can safely operate. The domain includes geographies, roadway types, speed range, weather, time of day, and state and local traffic laws and regulations. We describe the scenario and the ODD in which our simulated ADS-equipped vehicles safety will be assessed.

A. ODD Description

The vehicle in this ODD is moving along a single lane road and performing a subset of the Federal Test Procedure 75 (FTP75) drive cycle generated by UCEF as shown in Figure 5.

We focus on a subset of the FTP75 drive cycle [150s - 350s] where vehicle speed falls from 90km/h to 0km/h when an obstacle is detected.

We represent this subset on a [0s- 200s] time scale.

At t = T_control, the sensing component in UCEF triggers an alert that indicates an obstacle detection. The vehicle completely stops at t = T_actuation. For a given ODD there is an interval of time ΔT in which the vehicle performs full braking and completely stops when an obstacle is detected. ΔT can be expressed as follows:

$$\Delta T = T\_actuation - T\_control$$

The control component in UCEF generates a drive cycle based on the obstacle detection information it receives.

The Actuation Component (IGNITE-based) receives this drive cycle and simulates the vehicle response. We can compare both input drive cycle and output vehicle response in order to judge whether the vehicle has successfully performed the desired braking time. We assess ΔT once an obstacle is detected. Figure 6 shows vehicle velocity in response to input drive cycle generated by UCEF.

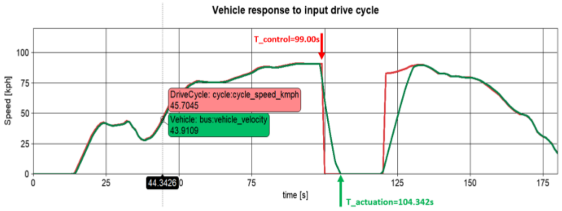


Fig. 6. IGNITE RPOST : Vehicle response to the input drive cycle: the vehicle completely stops 5.342s after the intended theoretical drive cycle control input. T_control and T_actuation are illustrated.

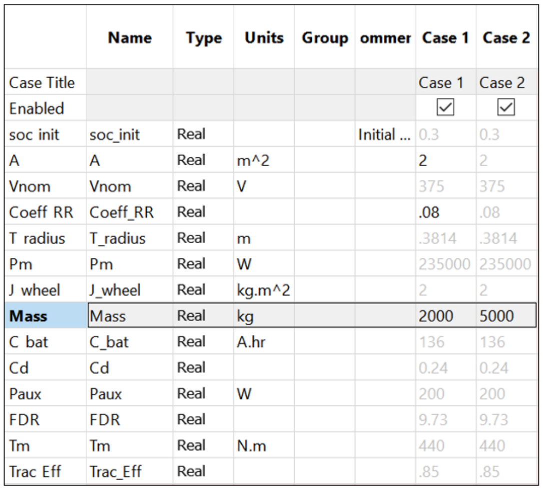
Halba, Khalid; Griffor, Edward; Roth, Thomas; Kamongi, Patrick. "Using Statistical Methods and Co-Simulation to Evaluate ADS-Equipped Vehicle Trustworthiness." Paper presented at Electric Vehicles International Conference (EV) 2019, Bucharest, RO. October 03, 2019 - October 04, 2019.

We have calculated for the default IGNITE parameters a Stopping Time equals to ΔT=5.342s between the drive cycle braking control message (T_control) and vehicle response (T_actuation). This Stopping Time is assessed within the described ODD to determine whether it falls within safe boundaries, i.e.: stopping before hitting the detected obstacle. A safe scenario verifies the following property:

$$0 < \Delta T < \Delta T\_Limit$$

ΔT_Limit is the time the vehicle is predicted to collide, and the Stopping Time will be beyond safe boundaries.

ΔT depends not only on the drive cycle, which is the result of decision support algorithms implemented in UCEF's control component, but it is also influenced by information collected by UCEF's sensing component and IGNITE's vehicle model parameters (vehicle mass for example); changing a single parameter may have a significant impact. Figure 7 is a list of vehicle parameters that could influence ΔT.



	Name	Type	Units	Group	ommer	Case 1	Case 2
Case Title						Case 1	Case 2
Enabled						☑	☑
soc init	soc_init	Real			Initial ...	0.3	0.3
A	A	Real	m^2			2	2
Vnom	Vnom	Real	V			375	375
Coeff RR	Coeff_RR	Real				.08	.08
T radius	T_radius	Real	m			.3814	.3814
Pm	Pm	Real	W			235000	235000
J wheel	J_wheel	Real	kg.m^2			2	2
Mass	Mass	Real	kg			2000	5000
C bat	C_bat	Real	A.hr			136	136
Cd	Cd	Real				0.24	0.24
Paux	Paux	Real	W			200	200
FDR	FDR	Real				9.73	9.73
Tm	Tm	Real	N.m			440	440
Trac Eff	Trac_Eff	Real				.85	.85

Fig. 7. A subset of IGNITE's vehicle model parameters. Different ODDs can be defined to assess a trustworthiness metric such as safety. We can define multiple cases each representing a run sequence, where a single or multiple parameters are altered. This figure shows two cases where vehicle mass was the altered parameter.

B. *Trustworthiness metric assessment using Fractional Factorial Experiment Design.*

One of the goals of the UCEF-based ADS-equipped vehicle testbed is to study ADS-equipped vehicles trustworthiness metrics. As in the NIST CPS Framework [18], trustworthiness comprises safety, security, privacy, resilience and reliability. This work has focused on illustrating the potential of co-simulation for assessing ADS-equipped vehicle safety measurement strategies. We have used the Fractional Factorial Experiment Design methodology to run multiple experiments while varying parameters of interests.

Many factors can impact the results of the experiments. Going forward we aim to determine the relative importance of these factors. The sheer number of these factors can present challenges. In this section we use the characteristics of the UCEF-based ADS-equipped vehicle testbed to study these parameters and to design our experiments, using the statistical Design of Experiments (DEX) methodology. Four factors were identified for assessing their influence on the output parameter ΔT (the time required for vehicle speed to go from 90km/h to 0km/h). The identified parameters are vehicle mass, tire rolling resistance coefficient, vehicle aerodynamics resistance surface, and wind speed.

The experiment design we adopt is a 2^k design that takes into consideration 2 levels per factor. This approach is best suited for exploratory experimentation purposes. The outcome of the 2^k factorial experiment helps in identifying the relative importance of factors and offers rapid insight into the interaction effects (Table I). The Fractional design is expressed as follows:

Design expression: $L^{(K-P)}$
L: number of levels of each factor investigated
K: number of factors investigated
P: size of the fraction of the full factorial to be eliminated
L^P: fraction of the full design L^K
M: number of experiments

TABLE I. LIST OF INPUT PARAMETERS CONSIDERED IN THE TWO-LEVEL FRACTIONAL FACTORIAL DESIGN, AND THE VALUES CHOSEN FOR THE TWO LEVELS FOR EACH VARIABLE.

Variable	Description	Low (-1)	High (+1)
X1	Vehicle Mass (Mass)	2000kg	5000kg
X2	Tire Rolling Resistance Coefficient	0.01	0.08
X3	Vehicle Aerodynamics Resistance Surface (Area)	$2m^2$	$5m^2$
X4	Wind Speed	0 mph	45 mph

TABLE II. ΔT IS THE OUTPUT PARAMETER (RESPONSE) MEASURED BASED ON VARIATIONS OF VEHICLE PROPERTIES AND WEATHER CONDITIONS.

Output parameter Y1 = ΔT = F (X1, X2, X3, X4)
Y1: Stopping Time

For a system with K=4 factors, L=2 levels of each factor, and P=1, the number of experiments in a Fractional 2^{K-P} will be $M = 2^{4-1} = 8$ experiments. The half factorial design would reduce M, the number of experiments by half. Table III describes the different sets of experiments with the fractional 2^{K-P} experiment design without replication.

We have explained in this section how the experiment design using statistical fractional factorial techniques can be used to discern which factors yields the most significant response on the output parameters of interest and as a result assess safety related to the specified ODD parameters. By assessing which vehicle characteristics and road conditions have the most influence on ΔT outcomes, we can understand which factors have the most significant impact on the stopping time.

Halba, Khalid; Griffor, Edward; Roth, Thomas; Kamongi, Patrick. "Using Statistical Methods and Co-Simulation to Evaluate ADS-Equipped Vehicle Trustworthiness." Paper presented at Electric Vehicles International Conference (EV) 2019, Bucharest, RO. October 03, 2019 - October 04, 2019.

TABLE III. FRACTIONAL FACTORIAL 2K-P (K=4, P=1, M= 8)

X1	X2	X3	X4	RunSeq
-1	-1	-1	-1	1
1	-1	-1	1	2
-1	1	-1	1	3
1	1	-1	-1	4
-1	-1	1	1	5
1	-1	1	-1	6
-1	1	1	-1	7
1	1	1	1	8

The DEX mean plot [19] is appropriate for analyzing data from a designated experiment, with respect to important factors, where the factors are at two or more levels. The plot shows mean values for two or more levels of each factor plotted by factor. The mean values of a single factor are connected by a straight line. For the given factor levels results shows that "Vehicle Mass" is by far the most important factor in influencing ΔT. "Tire Rolling Resistance" plays the next most critical role. The experimental parameters tested for "Wind speed" and "Vehicle Aerodynamics Resistance Surface" did not indicate a statistically significant effect on stopping time. The average Stopping Time ΔTavg across all experiments is equal to 5.44s. if ΔT_Limit = ΔTavg, we can evaluate safe braking boundaries just by reviewing the DEX mean plot (Figure 8). Raw results can be found in [20].

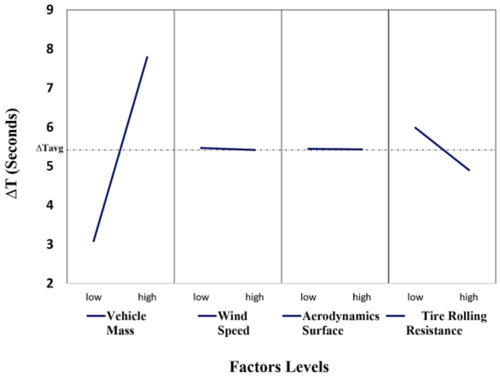


Fig. 8. The DEX mean plot: ΔT respose to 4 factors of interest.

V. CONCLUSIONS AND FUTURE WORK

In this work we have described the NIST UCEF-based-ADS-equipped vehicles testbed and its potential for co-simulation of ADS-equipped vehicle applications.

We have described implementation elements of the testbed and considered an ODD and a metric strategy for assessing safety, one component of trustworthiness.

Finally, we have described an experiment design methodology that can be used to assess trustworthiness metrics. Going forward, we intend to perform additional experiments to further study the safety and other trustworthiness metrics for ADS-equipped vehicles.

We will also design deep learning architectures to explore trustworthiness assessment techniques. For instance, we plan to leverage UCEF to study the safety of ADS functions and synthesize ground truths for training deep learning models. These deep learning models will be used for two purposes: to approximate the safety metric and forecast safety violations. Since many ADS functions are safety critical, we will design learning architectures that are interpretable and explainable.

REFERENCES

[1] Ricardo IGNITE Vehicle Simulator,
URL: https://software.ricardo.com/products/ignite.

[2] Vector CANoe,
URL:https://www.vector.com/int/en/products/products-a-z/software/canoe/.

[3] Vector CANalyzer,
URL:https://www.vector.com/int/en/products/products-a-z/software/canalyzer/.

[4] Feng, S., He, J. and Zhang, L., 2013. Modeling vehicle dynamics based on modelica. International Journal of Multimedia and Ubiquitous Engineering, 8(3), pp.307-318.

[5] Blochwitz, T., Otter, M., Arnold, M., Bausch, C., Elmqvist, H., Junghanns, A., Mauß, J., Monteiro, M., Neidhold, T., Neumerkel, D. and Olsson, H., 2011, June. The functional mockup interface for tool independent exchange of simulation models. In Proceedings of the 8th International Modelica Conference; March 20th-22nd; Technical Univeristy; Dresden; Germany (No. 063, pp. 105-114). Linköping University Electronic Press.

[6] Palmieri, M., Bernardeschi, C. and Masci, P., 2017, September. Co-simulation of semi-autonomous systems: the line follower robot case study. In International Conference on Software Engineering and Formal Methods (pp. 423-437). Springer, Cham.

[7] Bünte, T., Ho, L.M., Satzger, C. and Brembeck, J., 2014. Central vehicle dynamics control of the robotic research platform robomobil. ATZelektronik worldwide, 9(3), pp.58-64.

[8] Burns, M., Roth, T., Griffor, E., Boynton, P., Sztipanovits, J. and Neema, H., 2018, January. Universal CPS Environment for Federation (UCEF). In 2018 Winter Simulation Innovation Workshop.

[9] Anderson, V.L. and McLean, R.A., 2018. Design of experiments: a realistic approach.

[10] Khalid HALBA. Sensing and Control component : Sensing, Control, and Communication Federates:
URL https://github.com/KhalidHALBA-GR-NIST/UCEF-IGNITE.

[11] Khalid HALBA. Communication Module implementation: DriveCycle.FMU & an FTP75 DrveCycle Generator:
URL : https://github.com/KhalidHALBA-GR-NIST/FMU-IGNITE.

[12] Vertx Microservices Toolkit. URL: https://vertx.io/.

[13] VueJS, URL: https://vuejs.org/.

[14] Roth, T., Song, E., Burns, M., Neema, H., Emfinger, W. and Sztipanovits, J., 2017, June. Cyber-physical system development environment for energy applications. In ASME 2017 11th International Conference on Energy Sustainability.

[15] FTP75 Drive Cycle,
URL: https://www.dieselnet.com/standards/cycles/ftp75.php.

[16] Automated Driving Systems 2.0, A Vision for Safety, NHTSA, URL : https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf.

[17] Waymo Safety Report, On the Road to Fully Self-Driving, Waymo.
URL:https://storage.googleapis.com/sdc-prod/v1/safety-report/Safety%20Report%202018.pdf.

[18] Griffor, E.R., Greer, C., Wollman, D.A. and Burns, M.J., 2017. Framework for cyber-physical systems: Volume 1, overview(Special Publication (NIST SP)-1500-201).

[19] The DOE Mean Plot. Exploratory Data Analysis. URL : http://www.itl.nist.gov/div898/handbook/eda/section3/dexmeanp.htm

[20] Raw experiment results : URL : https://github.com/KhalidHALBA-GR-NIST/FMU-IGNITE/blob/master/dex%20mean%20plot.xlsx

40th AIVC - 8th TightVent & 6th venticool Conference, 2019

Residential Application of an Indoor Carbon Dioxide Metric

Andrew Persily, Brian J. Polidoro

National Institute of Standards and Technology
100 Bureau Drive, MS8600
Gaithersburg, MD 20899 USA
**Corresponding author: andyp@nist.gov*

ABSTRACT

Indoor carbon dioxide (CO_2) concentrations have been used for decades to evaluate indoor air quality (IAQ) and ventilation. However, many of these applications reflect a lack of understanding of the connection between indoor CO_2, ventilation rates and IAQ. In particular, a concentration of 1800 mg/m^3 (1000 ppm$_v$) has been used as a metric of IAQ and ventilation without an appreciation of its basis or application. After many years of trying to dissuade practitioners and researchers from using CO_2 as a metric of ventilation and IAQ, the first author developed an approach to determine CO_2 levels that can be used as more meaningful indicators. This approach is based on the fact that space types differ in their recommended or required ventilation rates, occupancy and other features that impact indoor CO_2 concentrations. Rather than employ a single CO_2 concentration for all spaces and occupancies, this alternative approach involves the estimation of space-specific CO_2 concentrations. The concept considers the steady-state CO_2 concentration that would be expected in a given space type based on its intended or expected ventilation rate per person, the time to achieve steady-state, the number of occupants as well as the rate at which they generate CO_2, and the occupancy schedule as it pertains to the likelihood that steady-state will be achieved.

This alternative approach was described in a previous AIVC conference paper, with sample calculations presented for several commercial and institutional building spaces. Those calculations yielded potential CO_2 concentration metrics, along with corresponding measurement times after full occupancy. Based on these analyses, it was stressed that reported CO_2 concentrations for comparison to these or other metrics need to be associated with a measurement time relative to the start of occupancy as well as information about the space in question and its occupancy. Since this previous work, an online calculator has been developed to allow users to perform these calculations, and that calculator is described here. In addition, this paper applies the approach to residential buildings, which are more challenging based on their varying configurations and the large fraction of time that occupants spend in bedrooms.

KEYWORDS

Building performance; carbon dioxide; indoor air quality; metrics; residential; ventilation

1 INTRODUCTION

Indoor air quality (IAQ) is characterized by the chemical and physical constituents of air that impact occupant health, comfort and productivity. The number of airborne contaminants in most indoor environments is quite large, and their impacts on building occupants are known for only a very small number of contaminants. The large number of contaminants, and their wide variation among and within buildings and over time, makes it extremely challenging to quantify IAQ, let alone to distinguish between good and bad IAQ based on a single metric. There have been efforts to define IAQ metrics, but none have been shown to capture the health and comfort

Persily, Andrew K.; Polidoro, Brian. "40th AIVC - 8th TightVent & 6th Venticool Conference, 2019 Residential Application of an Indoor Carbon Dioxide Metric." Paper presented at Conference 40th AIVC - 8th TightVent - 6th Venticool Conference in Ghent Belgium, Ghent, BE. October 15, 2019 - October 16, 2019.

impacts of IAQ very well or have become accepted in the field (Jackson et al., 2011; Hollick and Sangiovanni, 2000; Moschandreas et al., 2005; Teichman et al., 2015).

The indoor concentration of carbon dioxide (CO_2) has been widely promoted as a metric of IAQ and ventilation, in many cases without a clear explanation of what it is intended to characterize, or a description of its application or its limitations (Persily, 1997). Nevertheless, many practitioners use 1800 mg/m^3 (roughly 1000 ppm$_{v)}$ as a metric, often erroneously basing it on ASHRAE Standard 62.1 (ASHRAE, 2016a). However, that standard has not contained an indoor CO_2 limit for almost 30 years (Persily, 2015a). There have been many papers and presentations that have attempted to clarify the meaning of indoor CO_2 concentrations, some advocating that they not be used at all in IAQ and ventilation evaluations. However, these calls to stop poorly informed applications of indoor CO_2 are not succeeding. Instead, efforts to educate designers, practitioners and others need to continue, and this paper expands on a previously-described approach to using indoor CO_2 concentrations as a metric of ventilation rate per person that thoroughly considers the parameters that determine indoor CO_2 levels (Persily, 2018). That previous discussion presented the approach and outlined its application to a number of commercial and institutional building spaces. This paper expands the consideration to residential buildings and describes an on-line calculator that allows the estimation of indoor CO_2 concentrations in applying this approach.

2 BACKGROUND

Indoor CO_2 concentrations have been prominent in discussions of ventilation and IAQ since the 18th century (Klauss et al., 1970). Since that time, discussions of CO_2 in relation to IAQ and ventilation have evolved, focusing on the impacts of CO_2 concentrations on building occupants, how these concentrations relate to occupant perception of bioeffluents, the use of indoor CO_2 concentrations to estimate ventilation rates, and the control outdoor air ventilation rates based on indoor CO_2 concentrations (Persily, 2015b; Persily, 1997).

Indoor CO_2 concentrations are directly related to the outdoor air ventilation rates per person specified in standards, guidelines and building regulations (ASHRAE, 2016a; ASHRAE, 2016b; CEN, 2007b; CEN, 2009). These outdoor air requirements reflect research on the amount of ventilation needed to control odor associated with the byproducts of human metabolism, as well as other contaminants emitted by building materials and furnishings (Persily, 2015a). This research has found that about 7.5 L/s to 9 L/s per person of ventilation air dilutes body odor to levels judged to be acceptable by individuals entering a room from clean air, i.e., unadapted visitors. This research also supports 1800 mg/m^3 of CO_2 as a reflection of body odor acceptability perceived by unadapted visitors. Of course, there are many other important indoor air contaminants that are not associated with the number of occupants, and CO_2 concentration is not a good indicator of those contaminants.

Indoor CO_2 concentrations are typically well below values of interest based on health concerns (Persily, 2017). Some recent work has shown evidence of impacts on human performance, as well as other health impacts, at levels on the order of 1800 mg/m^3 (Azuma et al., 2018; Snow et al., 2019), while other studies have not shown performance impacts at similar concentrations. It is therefore premature to conclusively link CO_2 concentrations in this range with such occupant impacts until more research is done.

While indoor CO_2 concentrations are not meaningful indicators of overall IAQ, a previous paper describes the use of CO_2 as an indicator or metric of outdoor air ventilation rates per person (Persily, 2018). As discussed in that paper, indoor CO_2 concentrations depend primarily on the rate at which the occupants generate CO_2, the outdoor air ventilation rate of the space, the time since occupancy began, and the outdoor CO_2 concentration. For the purposes of these discussions, outdoor air ventilation refers to the total rate at which outdoor air enters the building or space of interest, including mechanical and natural ventilation as well as infiltration. The cited paper describes the single-zone mass balance theory to calculate indoor CO_2 concentrations from these parameters per the following equation:

$$C(t) = C(0)e^{-\frac{Q}{V}t} + C_{ss}\left(1 - e^{-\frac{Q}{V}t}\right), \tag{1}$$

where C is the CO_2 concentration in the space in mg/m^3, $C(0)$ is the indoor concentration at t = 0, t is time in hours, Q is the volumetric flow of air into the space from outdoors and from the space to the outdoors in m^3/h, and V is the volume of the space being considered in m^3. The steady-state CO_2 concentration C_{ss} is given by:

$$C_{ss} = C_{out} + {}^{G}/_{Q}, \tag{2}$$

where C_{out} is the outdoor CO_2 concentration and G is the CO_2 generation rate in the space in mg/h. Q, C_{out} and G are in general functions of time but are assumed constant in this analysis. Also, air density differences between indoors and out are being ignored by using the same value of Q for the airflow into and out of the space. Finally, this single zone formulation ignores concentration differences within and between building zones and CO_2 transport between zones and assumes there are no other indoor sources of CO_2 other than occupants.

Note that the indoor concentration will only get sufficiently close to steady-state if conditions, specifically Q and G, are constant for a long enough period of time. In particular, a constant value of G requires that the occupancy remain constant, but in many spaces occupancy will be too short or too variable for steady-state to be achieved. A convenient means of assessing whether steady-state is likely to be achieved is by comparing the duration of constant occupancy to the time constant of the space. The time constant is equal to the inverse of Q/V in Equation 1, i.e., the inverse of the air change rate, and the indoor concentration will be about 95 % of steady-state after three time constants. For example, for an air change rate of 1 h^{-1}, steady-state will exist after 3 hours. For an air change rate of 0.5 h^{-1}, it will take 6 hours.

Table 1: Calculated CO_2 concentrations

			CO_2 concentration above outdoors (mg/m^3)		
Space Type	**t_{metric} (h)**	**Time to steady-state (h)***	**Steady-state**	**1 h**	**t_{metric}**
Classroom (5 to 8 y)	2	1.4	1060	940	1040
Classroom (>9 y)	2	1.1	1580	1490	1580
Lecture classroom	1	0.9	1940	1870	1870
Restaurant	2	0.7	1871	1850	1870
Conference room	1	1.6	2526	2130	2130
Hotel/motel bedroom	6	4.5	1080	520	1060
Office space	2	5.9	985	390	630
Auditorium	1	0.6	2900	2880	2880
Lobby	1	0.6	4467	4430	4430
Retail/Sales	2	2.1	1546	1170	1450

* Time to achieve 95 % of steady-state CO_2 concentration, i.e., three time constants

In the previous work on this CO_2 metric concept, several space types were selected from the commercial/institutional building space types or "Occupancy Categories" in Table 6.2.2.1 of ASHRAE Standard 62.1 (ASHRAE, 2016a). For these spaces, shown in Table 1, CO_2 concentrations above outdoors were calculated at steady-state, after 1 h of occupancy, and at a time t_{metric}, which was selected as a time over which the particular space type may be expected to be fully occupied. That time is in the first column of Table 1, while the time to reach steady-state is in the second column. The assumed occupant densities, occupant characteristics and CO_2 generation rates are described in Persily (2018).

Based on the previous work, including the desire for a CO_2 metric to capture ventilation deficiencies and to be less sensitive to the timing of the concentration measurement, Table 2 summarizes potential CO_2 metric values for these spaces along with the corresponding measurement time. Given the transient nature of indoor CO_2 concentrations, it is critical that a concentration CO_2 metric be linked to a measurement time. Therefore, reported CO_2 concentrations relative to these and any other metrics need to include the time that has passed since the space reached full occupancy. Consideration of additional spaces and different input values would possibly yield other conclusions about potential metrics. Such analyses will be facilitated by the online tool described below.

Table 2: Potential CO_2 concentration metrics

Space Type	CO_2 concentration metric, above outdoors (mg/m^3)	Corresponding time (h after full occupancy)
Classroom (5 to 8 y)	1000	2
Classroom (>9 y)	1500	1
Lecture classroom	2000	1
Restaurant dining room	2000	1
Conference meeting room	2000	1
Hotel/motel bedroom	1000	6
Office space	600	2
Public assembly/Auditorium	3000	1
Public assembly/Lobby	4500	1
Retail/Sales	1500	2

3 CO_2-BASED VENTILATION METRIC FOR RESIDENTIAL SPACES

This paper extends the concepts discussed above to residential spaces, which can be challenging given the variations in dwelling and family size and in occupant characteristics, as well as the often unpredictable durations of occupancy relative to some commercial and institutional spaces. However, the many hours associated with sleep provide helpful options for these analyses in bedrooms. The approach taken is again to use Equation (1) to calculate the CO_2 concentrations for a given space based on assumptions about the CO_2 generation rates and ventilation rate of the space. In order to explore these dependencies for residential spaces, indoor CO_2 concentrations were calculated for the occupancies listed in Table 3, which describes 3 families: a baseline with 4 members (2 adults and 2 children), a larger family with 2 additional children, and a smaller family with 2 adults and no children. The sex, age, body mass and level of physical activity are described for each family, including the CO_2 generation rate in L/s for each person calculated using the methodology in Persily and de Jonge (2017), as well as the average CO_2 generation rate per person. These generation rates are presented for the whole house during non-sleep hours when occupants are assumed to be more active, and for bedrooms when occupants are sleeping. For occupant characteristics that differ from those considered here, the online tool described below is enables analysis of different occupancies. For each occupancy in Table 3, CO_2 concentrations were calculated for the following ventilation scenarios:

Whole house:

- Ventilation rate requirement from ASHRAE Standard 62.2 (ASHRAE, 2016b)
- Ventilation rate of 0.5 h^{-1}

Bedrooms:

- 62.2/Perfect Distribution: Bedroom ventilation rate is the Standard 62.2 rate divided by the number of house occupants, multiplied by the number of bedroom occupants
- 62.2/Uniform Distribution: Bedroom ventilation rate is the Standard 62.2 rate divided by the whole house floor area, multiplied by the bedroom floor area
- 0.5/Perfect Distribution: Bedroom ventilation rate is 0.5 h^{-1} times the house volume divided by the number of house occupants and then multiplied by the number of occupants in each bedroom
- 0.5/Uniform Distribution: Bedroom ventilation rate is 0.5 h^{-1} times the house volume divided by the whole house floor area, and then multiplied by the bedroom floor area
- 10 L/s per person/Perfect Distribution: Bedroom ventilation rate is 10 L/s multiplied by the number of bedroom occupants

Table 3: Occupancy Assumptions for CO_2 concentration calculations

Case	**Occupants** **(**age, body mass in kg, met level**)**	**CO_2 generation per person (L/s)**	**Average CO_2 generation per person (L/s)**
Baseline family of 4			

Whole house	1 male (40 y, 85 kg, 1.3 met);
	1 female (40 y, 75 kg, 1.3 met);
	1 male (6 y, 23 kg, 2 met);
	1 female (10 y, 40 kg, 1.7 met)
		0.0049
		0.0038
		0.0042
		0.0042
			0.0043

Master Bedroom	1 male (40 y, 85 kg, 1.3 met);
	1 female (40 y, 75 kg, 1.3 met);
		0.0037
		0.0029
			0.0033

Child Bedrooms	1 male (6 y, 23 kg, 2 met); 1 female (10 y, 40 kg, 1.7 met)	0.0021 0.0025	0.0023
Additional occupants in larger family of 6			
Whole house	1 male (8 y, 32 kg, 2 met); 1 female (4 y, 14 kg, 2 met)	0.0050 0.0031	0.0042*
Master Bedroom	No change		0.0033
Child Bedrooms	1 male (8 y, 23 kg, 2 met); 1 female (4 y, 40 kg, 1.7 met)		0.0022*
Smaller family of 2 (no children)			
Whole house	Only adults		0.0043
Master Bedroom	Only adults		0.0033

* Average CO_2 generation rate accounts for all 6 occupants in whole house and all 4 children in child bedrooms.

The Standard 62.2 whole house ventilation requirement Q_{tot} in L/s is calculated using Equation (4.1b) from the standard, i.e.,

$$Q_{tot} = 0.15A_f + 3.5(N_{br} + 1), \tag{2}$$

where A_f is the floor area in m^2 and N_{br} is the number of bedrooms. The value of Q_{tot} is used in this analysis without any of the adjustments allowed by Standard 62.2, such as the infiltration credit. The whole house air change rate of 0.5 h^{-1} is included as it is recommended in several international standards and guidelines (CEN, 2007a; Concannon, 2002).

For the bedroom cases, two idealized air distribution scenarios are applied to the Standard 62.2 and 0.5 h^{-1} whole house rates. In the first, Perfect Distribution, the whole house rate is divided by the number of occupants in the house. That normalized value is multipled by the number of occupants in each bedroom to determine the ventilation to each bedroom. Perfect Distibution may correspond to a ventlation system that supplies outdoor air directly to each bedroom based on the number of occupants. Under Uniform Distribution, the total ventilaiton rate is normalized by the floor area of the entire house, and the ventilation rate of each bedroom is that normalized rate multiplied by its floor area. Uniform distribution may correspond to a building ventilated by infiltration only, an exhaust-only ventilation system or a mechanical ventilation system that is integrated into a forced-air distribution system. The last bedroom ventilation rate, 10 L/s per person/Perfect Distribution, assumes 10 L/s of outdoor air is supplied for each person in each bedroom. That rate is based on recommendations in CEN (2007a) and (2009).

Table 4 presents the dimensions (ceiling heights and floor areas) for the houses considered.

Table 4: House and bedroom sizes for cases considered

	Ceiling height (m)	**House floor area (m^2)**	**Master bedroom floor area (m^2)**	**Child bedrooms floor area (m^2)**
Large House	2.44	250	30	20
Small house	2.74	200	20	15

Table 5 presents ventilation rates and calculated CO_2 concentrations (above outdoors). The second and third columns contain the outdoor air ventilation rate in L/s per person and h^{-1} for the whole house and bedroom cases. The fourth column is the time to reach a steady-state CO_2 concentration, i.e., three times the inverse of the air change rate. The table does not include t_{metric}, which as described earlier is a time over which a particular space may be expected to be fully occupied. Throughout these analyses, t_{metric} is 2 h for the whole house and 6 h for the bedrooms. The last three columns are the calculated CO_2 concentrations at steady-state, 1 h after occupancy and t_{metric}. The whole house air change rates based on Standard 62.2 are about 0.3 h^{-1} in all but the small house/baseline family case, in which it is about 0.5 h^{-1}. The bedroom air change rates cover a range of almost 10 to 1 for the different cases in the large house/baseline family and the small house/small family. In both of those occupancies, delivering 0.5 h^{-1} directly to the bedrooms under Perfect Distribution results in air change rates above 2 h^{-1} and well over 15 L/s per person. On the other hand, Uniform Distribution to the bedrooms results in less than 5 L/s per person in several cases. These air change rates impact the time required to achieve steady-state, which are 6 h or less for the bedrooms cases other than for 62.2/Uniform. For the whole house, the time to steady-state ranges from 3 h to more than 11 h.

The calculated CO_2 concentrations in Table 5 reflect the differences in ventilation and CO_2 generation for the specific occupancies. It is worth noting that the only steady-state bedroom concentrations greater than 1800 mg/m^3 (assuming an outdoor concentration of 700 mg/m^3) occur in the master bedroom for all 62.2/Uniform cases and for the small house/small family 0.5/Uniform case. Also, for cases with short time constants, 2 h or less, the three CO_2 concentrations are not very different from each other. Much larger differences are seen for larger time constants. In all whole house cases, the concentrations at t_{metric} are at least 200 mg/m^3 above outdoors, which is above the uncertainty typically associated with field measurements of indoor CO_2 concentrations. The bedroom concentrations at t_{metric} are typically even higher, except in the 0.5 h^{-1}/CBR/Perfect case, for which the steady-state concentration is less than 200 mg/m^3. The magnitude of these concentrations relative to typical measurement uncertainties supports the use of such calculated concentrations as a metric, although assuming constant occupancy in a whole house for 2 h could be questionable under some circumstances. It is worth noting that the range of CO_2 concentrations at t_{metric} for each occupancy range by at least 4 to 1 in the small house/baseline family case, to as much as almost 9 to 1 in the large house baseline/family case. These differences demonstrate the importance of considering the target ventilation rate in using calculated CO_2 concentrations as a metric. Also, for the same ventilation case, the differences in calculated CO_2 concentrations at t_{metric} for the different occupancies are not as large as those within the same occupancy for the different ventilation cases. However, they are large enough to be reliably measured and support the need to consider house size and occupancy when using calculated concentrations as a metric.

Table 5: Ventilation rates and calculated CO_2 concentrations for the residential cases

	Outdoor air ventilation		**Time to steady-state (h)**	**CO_2 Concentration above outdoors (mg/m^3)**		
Case*	**L/s per person**	**h^{-1}**		**Steady-state**	**1 h**	**t_{metric}**
Large House/Baseline family						
Whole house – 62.2	12.9	0.27	11.1	598	142	250
Whole house – 0.5 h^{-1}	23.8	0.50	6.0	324	127	205
62.2/ MBR/Perfect	12.9	1.13	2.7	461	312	461
62.2/ CBR/Perfect	12.9	0.85	3.5	322	184	320
62.2/MBR/Uniform	3.1	0.27	11.1	1922	456	1543
62.2/CBR/Uniform	4.1	0.27	11.1	1005	238	807
0.5 h^{-1}/MBR/Perfect	23.8	2.08	1.4	250	219	250
0.5 h^{-} CBR/Perfect	23.8	1.56	1.9	174	138	174
0.5 h^{-1}/MBR/Uniform	5.7	0.50	6.0	1041	409	989
0.5 h^{-1}/ CBR/Uniform	7.6	0.50	6.0	544	214	517
10 L/s per person/MBR/Perfect	10.0	0.88	3.4	594	347	591
10 L/s per person/CBR/Perfect	10.0	0.66	4.6	414	199	406
Large House/Large family						
Whole house – 62.2	9.8	0.31	9.8	775	205	356
Whole house – 0.5 h^{-1}	15.9	0.50	6.0	477	188	301
62.2/ MBR/Perfect	9.8	0.85	3.5	609	350	606
62.2/ CBR/Perfect	9.8	0.64	4.7	402	190	393
62.2/MBR/Uniform	3.5	0.31	9.8	1692	448	1425
62.2/ CBR/Uniform	4.7	0.31	9.8	837	221	704
0.5 h^{-1}/MBR/Perfect	15.9	1.39	2.2	375	281	375
0.5 h^{-} CBR/Perfect	15.9	1.04	2.9	247	160	246
0.5 h^{-1}/MBR/Uniform	5.7	0.50	6.0	1041	409	989
0.5 h^{-1}/ CBR/Uniform	7.6	0.50	6.0	514	202	489
10 L/s per person/MBR/Perfect	10.0	0.88	3.4	594	347	591
10 L/s per person/CBR/Perfect	10.0	0.66	4.6	392	189	384
Small House/Baseline family						
Whole house – 62.2	7.3	0.43	7.0	1061	369	610
Whole house – 0.5 h^{-1}	8.5	0.50	6.0	908	357	574
62.2/ MBR/Perfect	7.3	1.07	2.8	819	538	818
62.2/ CBR/Perfect	7.3	0.71	4.2	571	291	563
62.2/MBR/Uniform	2.9	0.43	7.0	2048	713	1891
62.2/ CBR/Uniform	4.4	0.43	7.0	952	331	879
0.5 h^{-1}/MBR/Perfect	8.5	1.25	2.4	701	500	701
0.5 h^{-} CBR/Perfect	8.5	0.83	3.6	489	276	485
0.5 h^{-1}/MBR/Uniform	3.4	0.50	6.0	1753	690	1666
0.5 h^{-1}/ CBR/Uniform	5.1	0.50	6.0	814	320	774
10 L/s per person/MBR/Perfect	10.0	1.48	2.0	594	458	594
10 L/s per person/CBR/Perfect	10.0	0.98	3.1	414	259	413
Small House/Small family						
Whole house – 62.2	11.0	0.32	9.2	700	194	334
Whole house – 0.5 h^{-1}	16.9	0.50	6.0	454	179	287
62.2/ MBR/Perfect	11.0	1.62	1.8	540	433	540
62.2/MBR/Uniform	2.2	0.32	9.2	2700	748	2315
0.5 h^{-1}/MBR/Perfect	16.9	2.50	1.2	351	322	351
0.5 h^{-1}/MBR/Uniform	3.4	0.50	6.0	1753	690	1666
10 L/s per person/MBR/Perfect	10.0	1.45	2.0	594	458	594

* MBR and CBR stand for master bedroom and child bedroom, respectively.

Table 6: Calculated CO_2 concentrations for the residential cases with 25 % ventilation rate reduction

	Outdoor air ventilation		**Time to steady-state (h)**	**CO_2 Concentration above outdoors (mg/m^3)**		
Case*	**L/s per person**	**h^{-1}**		**Steady-state**	**1 h**	**t_{metric}**
Large House/Baseline family						
Whole house – 62.2	10.5	0.22	13.6	**731**	145	261
Whole house – 0.5 h^{-1}	17.8	0.38	8.0	**431**	135	228
62.2/ MBR/Perfect	10.5	0.92	3.3	**564**	340	**562**
62.2/ CBR/Perfect	10.5	0.69	4.3	393	196	387
62.2/MBR/Uniform	2.5	0.22	13.6	**2350**	467	**1728**
62.2/CBR/Uniform	3.4	0.22	13.6	**1228**	244	**903**
0.5 h^{-1}/MBR/Perfect	17.8	1.56	1.9	333	263	333
0.5 h^{-} CBR/Perfect	17.8	1.17	2.6	232	160	232
0.5 h^{-1}/MBR/Uniform	4.3	0.38	8.0	**1387**	434	**1241**
0.5 h^{-1}/ CBR/Uniform	5.7	0.38	8.0	**725**	227	**649**
10 L/s per person/MBR/Perfect	7.5	0.66	4.6	**792**	381	**777**
10 L/s per person/CBR/Perfect	7.5	0.49	6.1	**552**	215	**523**
Large House/Large family						
Whole house – 62.2	8.2	0.26	11.6	**923**	210	372
Whole house – 0.5 h^{-1}	11.9	0.38	8.0	**636**	199	335
62.2/ MBR/Perfect	8.2	0.72	4.2	**725**	371	**716**
62.2/ CBR/Perfect	8.2	0.54	5.6	478	199	459
62.2/MBR/Uniform	2.9	0.26	11.6	**2015**	459	**1587**
62.2/ CBR/Uniform	3.9	0.26	11.6	**996**	227	785
0.5 h^{-1}/MBR/Perfect	11.9	1.04	2.9	**499**	323	**499**
0.5 h^{-} CBR/Perfect	11.9	0.78	3.8	329	178	326
0.5 h^{-1}/MBR/Uniform	4.3	0.38	8.0	**1387**	434	**1241**
0.5 h^{-1}/ CBR/Uniform	5.7	0.38	8.0	**686**	214	**614**
10 L/s per person/MBR/Perfect	7.5	0.66	4.6	**792**	381	**777**
10 L/s per person/CBR/Perfect	7.5	0.49	.1	**522**	203	**495**
Small House/Baseline family						
Whole house – 62.2	6.3	0.37	8.1	**1219**	379	640
Whole house – 0.5 h^{-1}	6.4	0.38	8.0	**1211**	379	639
62.2/ MBR/Perfect	6.3	0.93	3.2	**941**	570	**937**
62.2/ CBR/Perfect	6.3	0.62	4.8	656	303	640
62.2/MBR/Uniform	2.5	0.37	8.1	**2352**	732	**2101**
62.2/ CBR/Uniform	3.8	0.37	8.1	**1093**	340	976
0.5 h^{-1}/MBR/Perfect	6.4	0.94	3.2	**935**	569	**931**
0.5 h^{-} CBR/Perfect	6.4	0.63	4.8	**652**	303	**636**
0.5 h^{-1}/MBR/Uniform	2.5	0.38	8.0	**2337**	731	**2091**
0.5 h^{-1}/ CBR/Uniform	3.8	0.38	8.0	**1086**	340	**971**
10 L/s per person/MBR/Perfect	7.5	1.11	2.7	**792**	530	**791**
10 L/s per person/CBR/Perfect	7.5	0.74	4.1	**552**	288	**545**
Small House/Small family						
Whole house – 62.2	9.1	0.27	11.1	**843**	199	351
Whole house – 0.5 h^{-1}	12.7	0.38	8.0	**606**	189	319
62.2/ MBR/Perfect	9.1	1.35	2.2	**651**	482	**651**
62.2/MBR/Uniform	1.8	0.27	11.1	**3255**	768	**2608**
0.5 h^{-1}/MBR/Perfect	12.7	1.88	1.6	**467**	396	**467**
0.5 h^{-1}/MBR/Uniform	2.5	0.38	8.0	**2337**	731	**2091**
10 L/s per person/MBR/Perfect	7.5	1.11	2.7	**792**	530	**791**

* MBR and CBR stand for master bedroom and child bedroom, respectively.

Persily, Andrew K.; Polidoro, Brian. "40th AIVC - 8th TightVent & 6th Venticool Conference, 2019 Residential Application of an Indoor Carbon Dioxide Metric." Paper presented at Conference 40th AIVC - 8th TightVent - 6th Venticool Conference in Ghent Belgium, Ghent, BE. October 15, 2019 - October 16, 2019.

To the evaluate the usefulness of calculated CO_2 concentrations as ventilation metrics, the calculations presented in Table 5 were redone with the assumed ventilation rates reduced by 25 %. As in the commercial and institutional occupancies discussed in Persily (2018), these additional calculations were performed to assess how much the concentrations change at lower ventilation rates since a useful metric should capture such changes. The CO_2 concentrations for the reduced ventilation rates are shown in Table 6, as well as the outdoor ventilation rates for each case in L/s per person and h^{-1} and the times to reach steady-state. Values that increase by 100 mg/m^3 or more relative to the corresponding values in Table 5 are noted in bold font. The whole-house, reduced-ventilation concentrations at t_{metric} in Table 6 increase very little relative to the corresponding values in Table 5, often by 30 mg/m^3 or less, which is comparable with the measurement accuracy of many field measurements of CO_2 concentrations. This lack of increase is partly due to the long time constants of the whole-house cases, which allow little time for the concentration to increase after only 2 h. The increases in the bedroom concentrations are more significant, typically at least 100 mg/m^3 and often several hundred mg/m^3 higher for the reduced ventilation cases for a t_{metric} of 6 h. These larger increases for the bedroom support the use of a 6 h value of t_{metric} to capture ventilation deficiencies in bedrooms.

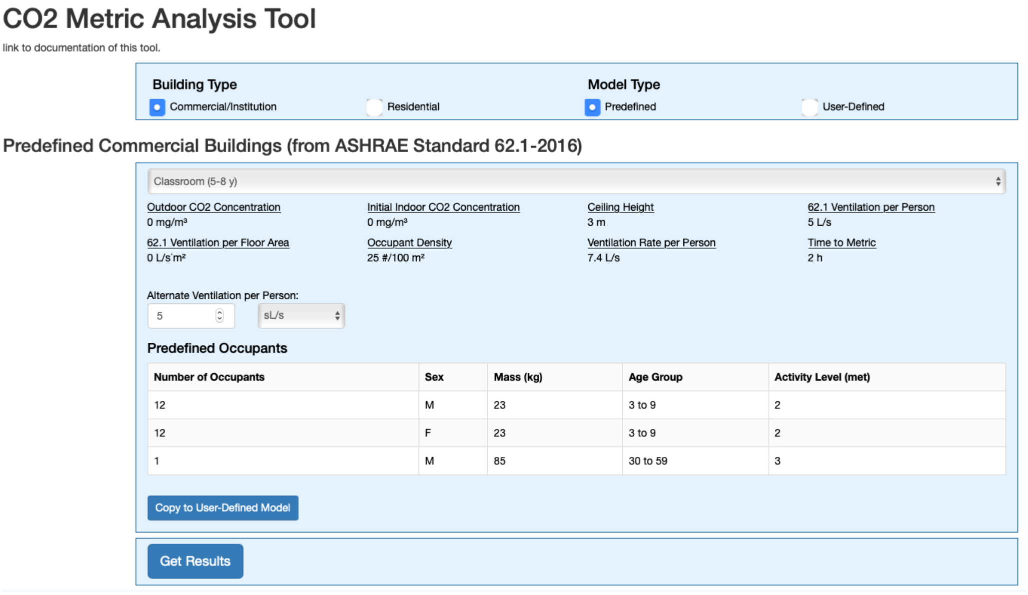
In contrast to the commercial and institutional occupancies discussed by (Persily, 2018), these residential cases are less constrained by space size, occupancy and ventilation, making it difficult to generalize these results to develop CO_2 metrics for residential buildings. Instead, the house, occupancy and air distribution approach need to be accounted for in developing a metric or reference point for evaluating the adequacy of the ventilation rate relative to a target value. The online tool discussed in the next section was developed to implement these concepts.

4 ON-LINE CO_2 METRIC CALCULATOR

In order to support application of the proposed CO_2 concentration metric, an online tool (available at https://pages.nist.gov/CONTAM-apps/webapps/CO2Tool/#/) has been developed. This tool allows the user to estimate indoor CO_2 concentrations in a ventilated space at steady-state, 1 h after occupancy and at a selected value of t_{metric}. These calculated concentrations can then be compared with measured concentrations in a building to evaluate whether the intended or required ventilation rate is actually being achieved. Such a building-specific metric or reference value is far better than using a single value such as 1800 mg/m^3.

Figure 1 shows the first screen encountered when using the tool, where one first selects to analyze a commercial/institutional building or a residential building. Depending on that selection, the user then enters the required inputs. For commercial/institutional buildings, the tool allows one to select from several of the commercial and institutional space types listed in ASHRAE Standard 62.1-2016, and to use the default values in that standard for outdoor ventilation requirements and occupant density, i.e., number of occupants per 100 m^2 of floor area. The tool makes assumptions about the occupants in each space, i.e., sex, body mass, age and activity level in met, needed to calculate the CO_2 generation rate in the space based on Persily and de Jonge (2017). However, these assumptions can be modified by selecting User-Defined Model Type, which brings up an alternative input screen.

The residential building inputs are shown in Figure 2. In this case, the user selects whether they are performing a whole building or bedroom analysis. If whole building is selected, the user can select the ventilation requirement based on Standard 62.2-2016 or enter a whole building air change rate in h^{-1}. If instead they chose to perform a bedroom analysis, they need to select the ventilation requirement from Standard 62.2 or enter a L/s per person ventilation rate. In either case, they also need to define the air distribution as Perfect or Uniform as described above. Under Perfect Distribution, they have the option of having some of the ventilation air bypass the bedrooms entirely, to account for supply vents in other portions of the house. If desired, an Alternate Ventilation per Person input can be input to enable comparison of the results to those obtained with the Primary ventilation rate.



CO2 Metric Analysis Tool

link to documentation of this tool.

Building Type

● Commercial/Institution ○ Residential

Model Type

● Predefined ○ User-Defined

Predefined Commercial Buildings (from ASHRAE Standard 62.1-2016)

Classroom (5-8 y)

Outdoor CO2 Concentration 0 mg/m³	Initial Indoor CO2 Concentration 0 mg/m³	Ceiling Height 3 m	62.1 Ventilation per Person 5 L/s
62.1 Ventilation per Floor Area 0 L/s˙m²	Occupant Density 25 #/100 m²	Ventilation Rate per Person 7.4 L/s	Time to Metric 2 h

Alternate Ventilation per Person:

5 sL/s

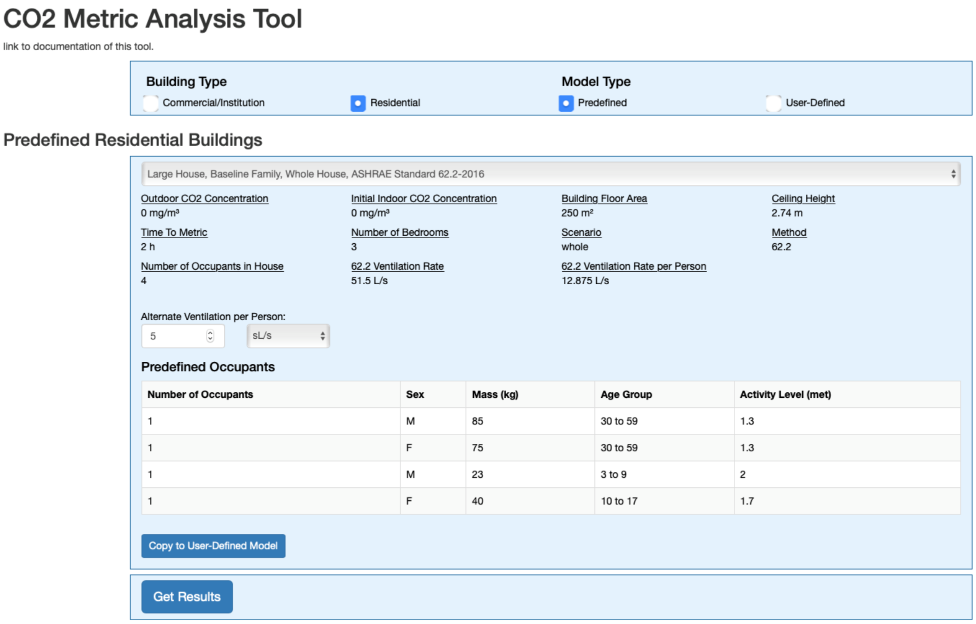
Predefined Occupants

Number of Occupants	Sex	Mass (kg)	Age Group	Activity Level (met)
12	M	23	3 to 9	2
12	F	23	3 to 9	2
1	M	85	30 to 59	3

Figure 1: CO_2 Metric Calculator Default Inputs Screen

Once the user has completed the inputs, they click Get Results on the bottom of the Inputs page. This action brings up the Results screen shown in Figure 3, which summarizes the inputs and displays a plot of the indoor CO_2 concentration versus time, along with concentration values at steady-state, t_{metric} and 1 h after occupancy for both the Primary and Alternate ventilation rates.

The tool is applied by comparing the calculated CO_2 concentrations to measured values, with a measured value that is higher serving as an indication that the actual ventilation rate is below the assumed or desired ventilation rate. For comparisons with a calculated whole house value, the measured CO_2 concentration should be a volume-weighted, whole house average based on the concentrations measured in each room. Since the calculation assumes constant occupancy, the measurement needs to occur while occupancy is constant, which can be limited in duration. Ideally, a constant occupancy period that lasts for t_{metric} occurs for the whole house, and the calculated value at that time are then used for the comparison. If constant occupancy does not last that long (default value of 2 h), the t_{metric} value in the calculator can be modified.



CO2 Metric Analysis Tool

link to documentation of this tool.

Building Type
○ Commercial/Institution ● Residential

Model Type
● Predefined ○ User-Defined

Predefined Residential Buildings

Large House, Baseline Family, Whole House, ASHRAE Standard 62.2-2016

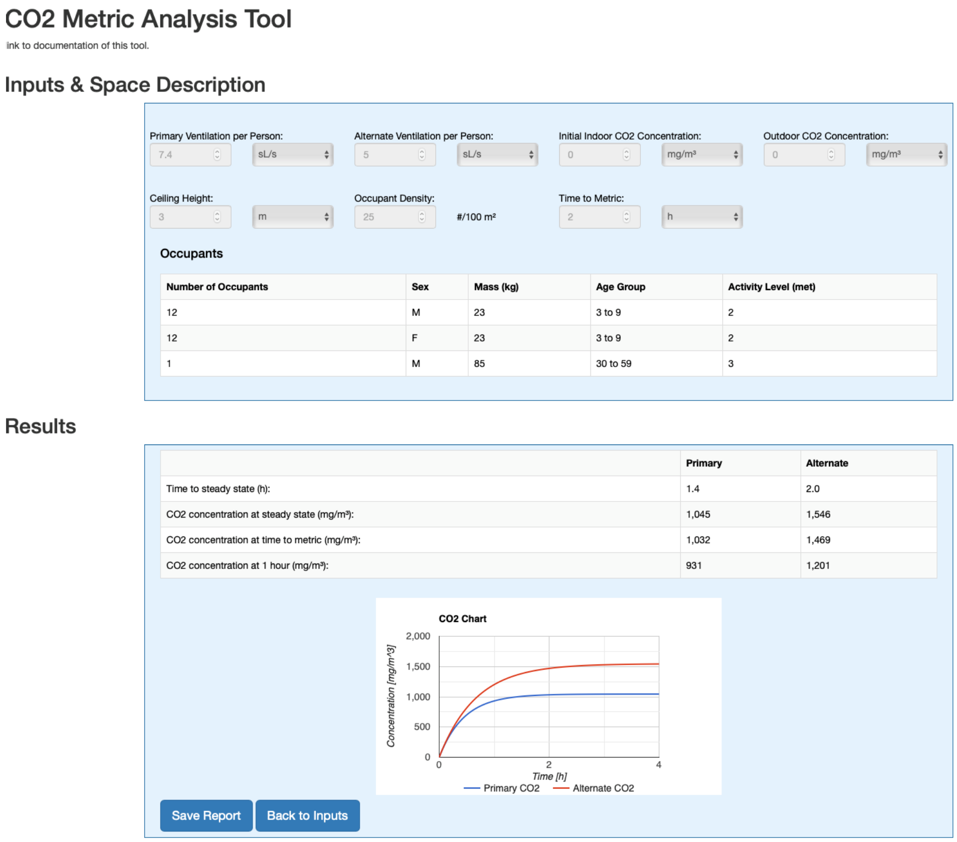
Outdoor CO2 Concentration 0 mg/m³	Initial Indoor CO2 Concentration 0 mg/m³	Building Floor Area 250 m²	Ceiling Height 2.74 m
Time To Metric 2 h	Number of Bedrooms 3	Scenario whole	Method 62.2
Number of Occupants in House 4	62.2 Ventilation Rate 51.5 L/s	62.2 Ventilation Rate per Person 12.875 L/s	

Alternate Ventilation per Person:
5 sL/s

Predefined Occupants

Number of Occupants	Sex	Mass (kg)	Age Group	Activity Level (met)
1	M	85	30 to 59	1.3
1	F	75	30 to 59	1.3
1	M	23	3 to 9	2
1	F	40	10 to 17	1.7

Figure 2: CO_2 Metric Calculator Residential Input Screen



CO2 Metric Analysis Tool

ink to documentation of this tool.

Inputs & Space Description

Primary Ventilation per Person: 7.4 sL/s

Alternate Ventilation per Person: 5 sL/s

Initial Indoor CO2 Concentration: 0 mg/m³

Outdoor CO2 Concentration: 0 mg/m³

Ceiling Height: 3 m

Occupant Density: 25 #/100 m²

Time to Metric: 2 h

Occupants

Number of Occupants	Sex	Mass (kg)	Age Group	Activity Level (met)
12	M	23	3 to 9	2
12	F	23	3 to 9	2
1	M	85	30 to 59	3

Results

	Primary	Alternate
Time to steady state (h):	1.4	2.0
CO2 concentration at steady state (mg/m³):	1,045	1,546
CO2 concentration at time to metric (mg/m³):	1,032	1,469
CO2 concentration at 1 hour (mg/m³):	931	1,201

Figure 3: CO_2 Metric Calculator Results Screen

Persily, Andrew K.; Polidoro, Brian. "40th AIVC - 8th TightVent & 6th Venticool Conference, 2019 Residential Application of an Indoor Carbon Dioxide Metric." Paper presented at Conference 40th AIVC - 8th TightVent - 6th Venticool Conference in Ghent Belgium, Ghent, BE. October 15, 2019 - October 16, 2019.

In the case of bedrooms, the calculated CO_2 concentrations can again be compared to measured values in the bedroom. Given the fairly stable bedroom occupancy during sleeping, that comparison should occur several hours after the bedroom is occupied for sleeping. The tool has a default value of t_{metric} for bedrooms of 6 h, which should work well for making the concentration comparisons. On making that comparison, a measured value that is higher serves as an indication that the actual ventilation rate is below the assumed or desired ventilation rate. Note that this comparison neglects the impact of interzone transport on the bedroom CO_2 concentration. Also, the calculation assumes that the CO_2 concentration starts at the outdoor level. However, the initial concentration in the bedroom may be higher than outdoors due to previous occupancy of the house, in which case the calculated concentration will be lower than it would if the actual initial concentration were considered. This situation would result in the calculated metric value being conservative, meaning it would lead to a conclusion that the ventilation rate is lower than it may actually be.

5 SUMMARY AND CONCLUSIONS

This paper expands on a previously-described approach to using indoor CO_2 concentration measurements as a metric for ventilation rates per person, which accounts for the ventilation requirements and occupancies of specific spaces. Calculations of CO_2 concentrations at steady state and other times, are presented for selected residential occupancies based on space-specific inputs of ventilation rate, space geometry and occupancy. Application of this CO_2 metric approach to residences requires one to report, at a minimum, the following: house or bedroom geometry (e.g. floor area and ceiling height), occupant characteristics, time at which full occupancy starts, time of CO_2 concentration measurement, and measured indoor and outdoor CO_2 concentrations. These measurements can then be compared with the values calculated with the online tool as an indication of whether the ventilation rate complies with the value in Standard 62.2 or other ventilation requirement of interest. As additional analyses are performed and the concept discussed with practitioners and researchers, it is anticipated that the approach will become more well-defined and more useful.

Note that all of the input values used in these calculations can be revised to examine the impact of other values on the resulting CO_2 concentrations. An online calculator has been developed to allow users to perform these additional calculations. Based on user feedback, the calculator will be revised in the future. One specific addition being considered is to enable Monte Carlo analyses to quantify the impact of uncertainties in the input values on the calculated CO_2 concentrations, as well as to identify the most important input values, using the methodology described in Jones et al. (2015).

6 ACKNOWLEDGEMENTS

The authors express their appreciation to Steven J. Emmerich and W. Stuart Dols of NIST and Benjamin Jones of the University of Nottingham for their helpful review comments.

7 REFERENCES

ASHRAE. 2016a. *ANSI/ASHRAE Standard 62.1-2016 Ventilation for Acceptable Indoor Air Quality*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

ASHRAE. 2016b. *ANSI/ASHRAE Standard 62.2-2016 Ventilation and Acceptable Indoor Air Quality in Low-Rise Residential Buildings*, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA.

Azuma, K., Kagi, N., Yanagi, U. and Osawa, H. 2018. Effects of low-level inhalation exposure to carbon dioxide in indoor environments: A short review on human health and psychomotor performance. *Environ Int*, 121, 51-56.

CEN. 2007a. *Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics*, Brussels, European Committee for Standardization.

CEN. 2007b. *Ventilation for buildings - Energy performance of buildings - Guidelines for inspection of ventilation systems*, Brussels, European Committee for Standardization.

CEN. 2009. *Ventilation for buildings - Determining performance criteria for residential ventilation systems*, Brussels, European Committee for Standardization.

Concannon, P. 2002. *Residential Ventilation*, Air Infiltration and Ventilation Centre, Coventry, Great Britain., Technical Note AIVC 57.

Hollick, H.H. and Sangiovanni, J.J. (2000) A Proposed Indoor Air Quality Metric for Estimation of the Combined Effects of Gaseous Contaminants on Human Health and Comfort, In: Nagda, N. L. (ed) *Air Quality and Comfort in Airliner Cabins, ASTM STP 1393*, West Conshohocken, PA, American Society for Testing and Materials, 76-98.

Jackson, M.C., Penn, R.L., Aldred, J.R., Zeliger, H.I., Cude, G.E., Neace, L.M., Kuhs, J.F. and Corsi, R.L. (2011) Comparison Of Metrics For Characterizing The Quality Of Indoor Air, *12th International Conference on Indoor Air Quality and Climate*, Austin, Texas.

Jones, B., Das, P., Chalabi, Z., Davies, M., Hamilton, I., Lowe, R., Mavrogianni, A., Robinson, D. and Taylor, J. 2015. Assessing uncertainty in housing stock infiltration rates and associated heat loss: English and UK case studies. *Building and Environment*, 92, 644-656.

Klauss, A.K., Tull, R.H., Roots, L.M. and Pfafflin, J.R. 1970. History of the Changing Concepts in Ventilation Requirements. *ASHRAE Journal*, 12, 51-55.

Moschandreas, D., Yoon, S. and Demirev, D. 2005. Validation of the Indoor Environmental Index and Its Ability to Assess In-Office Air Quality. *Indoor Air*, 15 (11), 874-877.

Persily, A. 2015a. Challenges in developing ventilation and indoor air quality standards: The story of ASHRAE Standard 62. *Building and Environment*, 91, 61-69.

Persily, A. (2017) Indoor Carbon Dioxide as Metric of Ventilation and iAQ: Yes or No or Maybe?, *AIVC 2017 Workshop on IAQ Metrics*, Brussels, Air Infiltration and Ventilation Centre.

Persily, A. (2018) Development of an Indoor Carbon Dioxide Metric, *39th AIVC Conference*, Antibes Juan-les-Pins, France, 791-800.

Persily, A.K. 1997. Evaluating Building IAQ and Ventilation with Indoor Carbon Dioxide. *ASHRAE Transactions*, 103 (2), 193-204.

Persily, A.K. (2015b) Indoor Carbon Dioxide Concentrations in Ventilation and Indoor Air Quality Standards, *36th AIVC Conference Effective Ventilation in High Performance Buildings*, Madrid, Spain, Air Infiltration and Ventilation Centre, 810-819.

Persily, A.K. and de Jonge, L. 2017. Carbon Dioxide Generation Rates of Building Occupants. *Indoor Air*, 27, 868-879.

Snow, S., Boyson, A.S., Paas, K.H.W., Gough, H., King, M.-F., Barlow, J., Noakes, C.J. and schraefel, m.c. 2019. Exploring the physiological, neurophysiological and cognitive performance effects of elevated carbon dioxide concentrations indoors. *Building and Environment*, 156, 243-252.

Teichman, K., Howard-Reed, C., Persily, A. and Emmerich, S. 2015. *Characterizing Indoor Air Quality Performance Using a Graphical Approach*, National Institute of Standards and Technology.

Augmenting Fiat Currency with an Integrated Managed Cryptocurrency

Peter Mell
National Institute of Standards and Technology
Gaithersburg MD, USA
peter.mell@nist.gov

***Abstract*—In this work, we investigate how the governance features of a managed currency (e.g., a fiat currency) can be built into a cryptocurrency in order to leverage potential benefits found in the use of blockchain technology and smart contracts. The resulting managed cryptocurrency can increase transparency and integrity, while potentially enabling the emergence of novel monetary instruments. It has similarities to cash in that it enables the general public to immediately transfer funds to a recipient without intermediary systems being involved. However, our system is account-based, unlike circulating bank notes that are self-contained. Our design would allow one to satisfy know your customer laws and be subject to law enforcement actions following legal due process (e.g., account freezing and fund seizure), while mitigating counterparty risk with checks and balances. Funds can thus be transferred only between approved and authenticated users. Our system has on-chain governance capabilities using smart contracts deployed on a dedicated, permissioned blockchain that has different sets of control mechanisms for who can read data, write data, and publish blocks. To enable the governance features, only authorized identity proofed entities can submit transactions. To enable privacy, only the block publishers can read the blockchain; the publishers maintain dedicated nodes that provide access controlled partial visibility of the blockchain data. Being permissioned, we can use a simple consensus protocol with no transaction fees. A separate security layer prevents denial of service and a balance of power mechanism prevents any small group of entities from having undue control. While permissioned, we ensure that no one entity controls the blockchain data or block publishing capability through a voting system with publicly visible election outcomes.**

***Index Terms*—Blockchain, Cryptocurrency, Digital Cash, Fiat Currency, Smart Contract**

I. Introduction

Bitcoin is a protocol for a permissionless distributed ledger that was designed to provide non-reversible transactions with direct account-to-account fund transfers where no third party needs to be trusted [1]. It leveraged blockchain technology to enable a form of non-sovereign digital currency that was previously not possible. It and subsequent cryptocurrencies introduced smart contracts and new kinds of decentralized governance models that have significant organizational and political implications (e.g., having no relationship with any government). With respect to these systems, [2] points out that cryptocurrencies can enable users to remain anonymous, can have permissionless access, and thus usually do not support know your customer (KYC) and anti-money laundering (AML) laws at the protocol level by design. In this work, we investigate how to leverage some of the novel benefits provided by blockchain technology and smart contracts to enable a new form of managed cryptocurrency that has built-in support for KYC and AML laws with system governance mechanisms along with a balance of power structure. Note that we are not suggesting that such a cryptocurrency should necessarily be issued, as that decision involves policy and economic factors outside of the scope of this work. Instead, we are proposing a technical architecture that could lead towards the technical ability to do so.

We investigate how the governance features of a managed currency (e.g., a fiat currency) can be built into a cryptocurrency in order to leverage potential benefits found in the use of blockchain technology and smart contracts. It is designed to be compatible with and augment a partner managed currency, the users being able to freely exchange one for the other. The resulting managed cryptocurrency can increase transparency and integrity, while potentially enabling the emergence of novel monetary instruments. It has similarities to cash in that it enables the general public to immediately transfer funds to a recipient without intermediary systems being involved and the associated counterparty risks (a single transaction to the system transfers funds). This is accomplished through a distributed multi-party managed cryptocurrency system providing guarantees similar to Bitcoin style cryptocurrencies. However unlike circulating bank notes, our system is account-based and all recipients are identity proofed and authorized. Our design thus supports the satisfaction of KYC and AML laws at the protocol level. Entities distinct from the platform and currency managers can register as identity providers, ensuring fund transfers only to identity proofed and authenticated recipients while maintaining openness to the private sector and competition. Accounts would also be subject to law enforcement actions following legal due process to include the freezing of accounts and fund seizure.

Our system has on-chain governance capabilities using smart contracts deployed on a dedicated, permissioned blockchain that has different sets of control mechanisms for who can read data, write data, and publish blocks. To enable the cryptocurrency to have built-in governance roles along with KYC/AML checks, only authorized identity proofed entities can submit transactions. To support user privacy features, only the miners (referred to henceforth as validators) can read

Mell, Peter. "Augmenting Fiat Currency with an Integrated Managed Cryptocurrency." Paper presented at The Fourteenth International Conference on Software Engineering Advances (ICSEA 2019), Valencia, ES. November 24, 2019 - November 28, 2019.

the blockchain. Validators then maintain dedicated nodes that provide access controlled partial visibility of blockchain data to users (e.g., their account balance, transaction history, and system management transactions). Being permissioned, we can use a lightweight consensus protocol. The protocol could be as simple as the dirty round robin used in Multichain [3]. The use of a security layer that prevents denial of service attacks (which works since all accounts must be pre-authorized and can easily be filtered) can enable a no transaction fee system where the validators are paid by the currency issuer to maintain the currency. Lastly, the architecture contains a balance of power mechanism to prevent any small group of entities from having undue control over the blockchain data or publication of new blocks. While it is a permissioned system, there is not a single entity that decides which accounts can publish blocks. Instead, the existing group of validators vote to determine changes to validator eligibility, with the outcomes being made publicly visible. No one entity controls the blockchain data or block publishing capability.

We implemented our architecture using smart contracts written with the Solidity programming language and made the code open source under a public domain license. It has functions for fund tracking, fiat-to-cryptocurrency fund conversion, transaction logging, account creation, voting scenarios, a bootstrapping mode, role assignment, and the ability of accounts to take special actions given their roles (e.g., law enforcement account freezing and central bank fund creation). A set of initial parameters are used to bootstrap the initial governance options but afterwards a voting system is used for multiple accounts with various roles to collectively manage different aspects of the cryptocurrency.

Research economists seem divided about the effectiveness of central bank issued cryptocurrencies. Some, like in [4], point out the potential economic viability of such assets. Others are more critical: [2] for instance states that there is a 'non-case' for a central bank cryptocurrency; their rationale for this was based on perceived immutable features of cryptocurrencies that would make them useless as alternatives to digital currency. In this work, we want to show that these features, considered immutable, can be altered through changing technical fundamentals about how a cryptocurrency blockchain works; this could enable sovereign cryptocurrencies with fiat currency style governance. Non-sovereign cryptocurrencies started the discussion on how to use blockchain to make the global financial system more stable and distributed; we hope that our work on sovereign cryptocurrencies will further facilitate that discussion.

The remainder of this paper is organized as follows. Section II discusses the foundational technology that underlies the design of our cryptocurrency platform. Section III presents our cryptocurrency architecture while section IV explains the account roles within that architecture. Section V discusses how to instantiate our cryptocurrency to integrate it with a fiat currency. Section VI analyzes the security model of our approach and Section VII discusses our implementation. Section VIII summarizes related efforts and section IX concludes.

II. Foundational Technology

Our managed cryptocurrency leverages existing approaches and borrows concepts from other technologies. This includes the cryptocurrency role system, on-chain governance of validator nodes, on-chain voting, and the decoupling of the validation and execution of transactions.

Our managed currency relies upon each cryptocurrency account being assigned a set of roles; these roles enable the management and use of the currency. Assigning roles to cryptocurrency accounts was initially introduced in [5].

On-chain governance is used to manage the set of validators through the assignment of 'validator' roles to accounts, those allowed to participate in a consensus algorithm to publish blocks. This concept of on-chain validator governance can be found in the Proof of Authority (PoA) consensus model. Here, block creation is distributed among different allowed nodes over time while offering Byzantine fault tolerance. PoA with smart contract based validator governance is implemented in the Ethereum client Parity (through the Aura consensus algorithm [6]) and by POA Network [7]. Another example of a PoA implementation can be found in the Microsoft Azure Blockchain system [8].

To manage the set of validators as well as other system functions, our managed cryptocurrency smart contracts must implement voting mechanisms which add or remove roles, as well as to approve or disapprove system security actions such as fund transfer reversals. Various standards and projects provide blockchain technology for decentralized voting [9]; the Ethereum standard EIP-1202 [10] offers for example an interface for implementing voting within smart contracts. As another example, the open-source project Aragon [11], built on Ethereum, allows token holders to cast a vote on protocol upgrades by signing a specific transaction. The Delegated Proof of Stake (DPoS) consensus algorithm, used for instance in BitShares [12], is another illustration of an on-chain voting structure. In BitShares, users vote by staking tokens into another account (called 'delegate'); the delegate account is then allowed to execute certain actions on behalf of its stakeholders (such as producing blocks and voting on protocol upgrades).

Lastly, our cryptocurrency introduces the concept of a 'Security Gateway'. A list of gateways linked to their associated validators, maintained at the smart contract level, are charged with pre-processing incoming transactions. This decouples the execution and validation of transactions. The Hyperledger Fabric, a permissioned blockchain, also has a similar decoupling [13]. Note that any mention of commercial products in this paper is for information only; it does not imply recommendation or endorsement.

III. Cryptocurrency Architecture

A cryptocurrency platform providing the benefits described in Section I and leveraging technology from Section II can be built using the following architecture. Figure 1 shows the overall architecture from the perspective of a single validator.

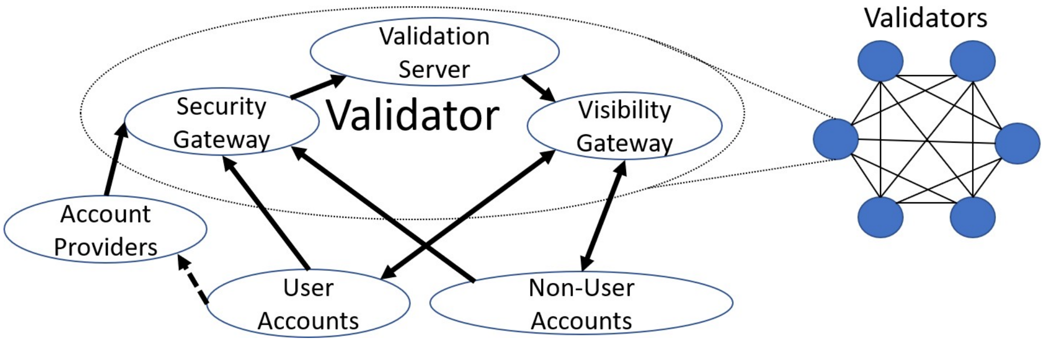


Fig. 1. Interactions of an Individual Validator

A. Platform Architecture

The architecture requires a permissioned smart contract cryptocurrency platform, such as PoA-based Ethereum. It should be configured to not charge transaction fees or gas for sending transactions to the smart contracts. The native cryptocurrency mechanisms of the platform will not be used and instead the cryptocurrency will be stored within the smart contracts (similar to token based ERC-20 [14] compliant smart contract currencies but running on a dedicated platform). Without gas and transaction fees, validators will be rewarded either off-chain or will participate through being inherently motivated to support the cryptocurrency. This is tractable because our use of a lightweight consensus model makes the execution of a validator node less expensive (it is done this way currently by other permissioned blockchain platforms such as Hyperledger Fabric).

The set of smart contracts will be fixed to a small set used to maintain the cryptocurrency. Being a permissioned system, the block publishing software must determine which validators are allowed to participate in the publication of new blocks. This set of permitted validators is managed at the smart contract layer and can be retrieved from the blockchain. In this way our architecture marries what are usually isolated governance layers, the protocol layer which manages the validator node permissions and the smart contract layer which executes code on behalf of system users.

B. Security Gateway

Also managed at the smart contract layer is the list of gateways that each validator maintains to accept proposed transactions from the users of the system. These security gateways pre-process incoming transactions to ascertain their validity. Transactions must be properly formatted and are only accepted from accounts that have roles. Security gateways also keep track of the rate of transactions issued by each account. Accounts with an unusually high rate can be throttled as a form of denial of service protection and to prevent any particular account from taking too large a percentage of system resources. It is possible to white list accounts that have a valid reason to issue a high throughput of transactions.

C. Visibility Gateway

A final platform level resource managed and made visible at the smart contract layer is the set of visibility gateways. Each validator independently maintains a set of such gateways to provide controlled blockchain read access for the account holders. The read access capabilities will be encoded as smart contract view functions (a view function is one that provides read only access and is highly efficient as it is executed locally and not propagated among validators nor included within a block like a normal transaction). However, unlike with typical view function responses, for security reasons (discussed in section VI) the responses will be signed by the associated validator account. The full blockchain is kept private by the validators and user read access is only available through the visibility gateways.

D. Smart Contracts

The smart contract layer, besides managing the authorized platform level resources listed previously, implements the managed cryptocurrency. The smart contracts maintain the list of authorized accounts, the roles granted to each account with associated features, and the balance in each account. We use the account/balance model as opposed to the unspent transaction output (UTXO) model (e.g., in Bitcoin) to avoid the unnecessary complexity (found in [5]) of having to label each unspent transaction with roles. The roles define a set of permissions that enable certain accounts to manage the cryptocurrency and are discussed in the following section.

E. Digital Wallets

Identity proofing results in the participants' identifiers being added to a global on-chain registry controlled by the account providers. As described in [15], the identifiers can be held in custodial, semi-custodial, or non-custodial digital wallets that can be integrated into existing applications, browsers, and operating systems.

IV. Managed Cryptocurrency Roles

The management features and integration with an associated fiat currency are enabled through accounts with various roles. This account and role capability is instantiated on top of the previously described platform and implemented within the fixed set of smart contracts. Section V will describe how these roles can be used in real world systems.

A. Platform Managers

An account with the platform manager role sets the policy for the cryptocurrency system and creates accounts and assigns them non-user roles. Policy can be set to be permanent, temporary, or have a timed expiration. Permanent policies cannot be changed once set (assuming the integrity of the blockchain itself is not compromised). They may be used to instantiate a particular architecture that the cryptocurrency will adopt. Alternately, they may be used to provide confidence to the user base that certain features or settings are guaranteed even though the cryptocurrency is managed by a set of privileged entities. Temporary policies can be changed at any time by a currency manager. Timed expiration policies are considered permanent until a published time at which they become temporary. The system may be set up with only one platform manager, a group of accounts that must vote to make changes, or a hierarchical system where higher priority managers can override policies from lower level managers (as in [5]). This latter design can be used as a security feature in case a currency manager account was compromised; higher priority accounts whose keys are stored in physical vaults could be used to override the compromised account and restore the system.

The policies available to be set can include enabling/disabling features within other roles, setting blockchain parameters such as the size and frequency of blocks, adjusting any fees charged (if any), and setting parameters on how voting will be performed (since the system requires groups to vote to perform certain actions).

During the bootstrapping phase for the cryptocurrency, within some fixed number of blocks, the platform manager defines the initial set of validators. Once the bootstrapping phase is over, the accounts with the platform manager role may not modify the validator roles (thus limiting their authority and creating a balance of power).

B. Account Providers

An account with an account provider role has been authorized by the platform manager(s) to manage user accounts. They identity proof users off-chain, receive a list of the users not yet authorized accounts, and add the user role to those accounts to authorize them. It is important that the users demonstrate ownership of each provided account through proving possession of the associated private key. Each account provider then keeps an internal record of which users are associated with which accounts; this record is not published or shared. This allows KYC and AML laws to be supported at the account provider level (rather than at the entire platform level), which may enhance security and user privacy.

C. System Security

An account with the system security role has the ability to control other accounts for system security purposes. Such accounts can freeze and unfreeze other accounts. They can also move funds between accounts to confiscate funds or reverse transactions. In the latter case we note that the relevant accounts simply need to be debited and credited funds due to our system being account-based (as opposed to following Bitcoin's UTXO model). We also note that since all accounts are identity proofed, system security actions can take place off-chain using existing legal frameworks.

To limit unauthorized actions, policy can be set by the platform manager requiring an on-chain voting mechanism for certain system security transactions. In addition, the platform manager(s) can limit or disable any of the powers of the system security role through policy settings.

D. Users

An account with the user role is one that can be used to receive, store, and send value in the form of tokens maintained by the smart contracts. A single user may have multiple accounts and may use multiple account providers to do so (note that every account must be identity proofed by an account provider).

Each account is labelled within the smart contract with its associated public key. A user maintains use of an account through possession of the associated private key (possibly stored on a hardware token for greater security). If a user loses a private key or suspects that their private key has been stolen, they need a way to retake possession of the account. This is accomplished by swapping out the account's original public key with a new one within the smart contract. When creating their account, users can choose what method they prefer to enable this action; there are at least three options. They can trust their account provider to do this for them and simply re-identity proof to their account provider. They could authorize a set of other accounts to validate the public key swap (using accounts they own or accounts of trusted individuals). Or they could require the involvement of system security along with their account provider, necessitating re-identity proofing with both entities.

E. Currency Managers

An account with the currency manager role has the ability to control the money supply through direct actions or ongoing policy. This includes fund creation, deletion, and the provision of interest. The currency manager accounts vote to set monetary policy or initiate an action (for example the creation of funds to be lent to other entities).

F. Validators

An account with the validator role is an account that represents an authorized block publisher. Validator accounts vote to add/remove the validator role to/from other accounts. Other than block publishing, the validators manage their respective security and visibility gateways. On their visibility gateways, they make visible all cryptocurrency management transactions to provide full transparency to all users.

Each validator account posts on the smart contract the Internet Protocol (IP) addresses of their security and visibility gateways. The visibility gateways then make the security and visibility gateway addresses visible to all users and the validation server addresses visible to other accounts with the validator role. This latter publication facilitates the peer-to-peer permissioned networking between validators used for transaction propagation and block publication.

Each validator account publishes on the blockchain a publicly visible special public key associated with the signed responses from its visibility gateways. This key is different from the public key for the validator account itself. It also publicly publishes contact information (e.g., an email address) for reporting any problems. This is essential for security reasons discussed in Section VI.

V. Integration with Fiat Currencies

The architecture presented in sections III and IV is designed to be integrated with a fiat currency and traditional bank deposits. A government administration could instantiate the cryptocurrency and act as the platform manager. The directors of the government's independent central bank could act as the currency managers. The government law enforcement agencies could act in the system security role. This creates a balance of power where no one organization 'controls' the blockchain. To further promote this, government entities separate from the administration can act as the validators (e.g., a set of states). The national standards body can define the specification for the supporting cryptocurrency software and independent laboratories can test compliance of that software. Note that multiple developers should be used, especially for the validator software, for security purposes and the code should be developed open source and made available publicly. This way a single developer cannot maliciously or unintentionally violate the specification and enable non-protocol compliant blocks to be published and accepted.

Financial institutions (e.g., commercial banks, cryptocurrency exchanges, and other fintech companies) could be made account providers, among other entities, since they already must identity proof their customers. They would keep their mapping of identity proofed users to account numbers private and only reveal select information to fulfill a court order (thus supporting KYC and AML laws while still maintaining user privacy). They can modify their banking software to simultaneously show users their bank deposit balances and cryptocurrency balances (since they established each user's accounts). The financial institution itself would not have access to the user's cryptocurrency balance and transactions but their banking application, on behalf of the user, could retrieve this information from the visibility gateways. These applications could then enable the conversion of bank deposits to cryptocurrency and vice versa (also often referred as on-ramp/off-ramp). The application could send cryptocurrency to the financial institution and have the institution deposit traditional money into the user's bank accounts (and vice versa). The application could also transfer funds between the user's different cryptocurrency accounts using a bank owned account as an intermediary to hide any linkage between the user's accounts from appearing on the blockchain. Note that the financial institution obtains cryptocurrency through its existing fiat accounts with the central bank: the institution sends the central bank fiat currency and the central bank sends it cryptocurrency. If users are allowed to interact directly with the central bank, users can perform this operation themselves.

The central bank, as the currency manager, unifies the fiat and cryptocurrency by enabling the exchange between both. The cryptocurrency could be maintained as a separate line item on the central bank balance sheet. The central bank can create and destroy both currencies and thus can implement a cryptocurrency monetary policy in a similar fashion as when managing solely its fiat currency. Note that offering two forms of currency, with different characteristics and risk profiles, can have significant economic implications that are out of scope for this paper.

VI. Security Analysis

In this section we analyze the security and functionality provided by our architecture using the three traditional computer security pillars of confidentiality, integrity, and availability.

A. *Confidentiality*

Our architecture provides accounts/transactions that are pseudonymous for the validators and confidential to the rest of the users. We note that there is a possibility of user confidentiality being lifted when necessary to support KYC and AML laws (e.g., through a court order for validators to reveal transactions and the respective account providers to divulge account ownership). Users choose which account provider they trust to know which accounts they own. The account provider keeps this private unless required to reveal it. Furthermore, the account provider can distribute user funds between the user's accounts such that there is no linkage on the blockchain between the multiple accounts from the same user.

The transactions on the blockchain are kept private and only shared within the set of validators. The visibility gateways only reveal blockchain transactions to the parties involved in those transactions. A downside of this is that it would appear then that accounts with the platform manager, system security, and currency manager roles can issue transactions without oversight. However, the validators are independent entities that make these transactions publicly visible through their visibility gateways. This offers transparency for all management transactions but confidentiality for user fund transfers.

The IP addresses of the validating servers are stored on the blockchain but the visibility gateways make this information visible only to the validator accounts. Thus, the validation servers themselves are kept confidential. If this information was leaked, a single transaction could be used to update a revealed server to a new IP address (to discourage denial of service (DoS) attacks). The security gateways and visibility gateways' addresses are made public. The visibility gateways operate independently with only a copy of the blockchain and thus can be replicated at scale to counter possible DoS attacks.

Likewise, a validator may have multiple security gateways to provide load balancing and DoS protection.

B. Integrity

The use of a group of independent validators ensures the integrity of the newly published blocks. No validator will accept a block from another validator that does not follow the established protocol. Each block and the transactions therein must be of the proper form and have the necessary digital signatures. As is the norm with blockchains, each block contains a hash of the previous block to enable detection of any changes with previously published blocks. However unlike public blockchain systems, the users themselves cannot verify the retained integrity of the blockchain and so another mechanism must exist to hold individual validators accountable.

For this, user software will query multiple visibility gateways when retrieving user blockchain data. Any discrepancy between the visibility gateways owned by different validators reveals a problem with one of the validators. An exception is for very recent transactions that have not yet been posted by all validators and thus there should thus be an agreed-upon time delay before taking any action. In the case of a discrepancy, an event transaction is triggered and sent to all validators to describe the discrepancy. As long as at least one validator is honest, the discrepancy will be published. Note that this is considered a 'management' type transaction and thus is made publicly visible to all users. Note that since the visibility gateways sign their responses, the user can prove that multiple visibility gateways provided different answers. As long as a majority of the validators remain honest, the honest ones can vote out any validators that provide incorrect results.

If a set of the validators decide to overtly violate the cryptocurrency protocol (e.g., to change a permanent policy or take control away from the platform or currency managers), this will fork the blockchain as happens with other cryptocurrency systems. The non-violating validators would inform participating parties using off-chain methods and, like other cryptocurrency forks, the resolution would take place off-chain. Given that the cryptocurrency will be tied to a fiat currency, investigations and legal action may be taken against the violating validators. Note that only if 100 percent of the validators collude can they make changes without being noticed. Also, note that our cryptocurrency leverages the fact that it is a sovereign currency, existing within an off-chain legal framework.

C. Availability

There are two types of availability that need to be considered: the availability of the cryptocurrency system as a whole and the availability of a particular account to conduct transactions.

The cryptocurrency platform itself has robust availability because it is a distributed system with no central point of failure. Many of the validation servers may fail, even the majority of them, and the system can still continue to function. Note that individual validation servers can be run efficiently due to the use of a lightweight consensus algorithm, permitted by the permissioned blockchain configuration. Also, each security and visibility gateway can be implemented as a cluster of servers to reduce susceptibility to DoS attacks and individual server failures.

Individual accounts are not dependent upon a particular validator and user applications should issue transactions to multiple systems simultaneously (this includes both write transactions to the security gateways and read transactions to the visibility gateways). Using multiple security gateways ensures that no validator could decide to unilaterally block a particular account (note that account ownership is pseudonymous to the validators making this less likely). Using multiple visibility gateways, as discussed above, addresses integrity concerns.

VII. Implementation

We implemented our managed cryptocurrency as smart contracts using the Solidity programming language. It is available as open source software on Github at https://github.com/usnistgov/managed_token under a public domain license. In this proof-of-concept prototype, we implemented the core functions of the managed cryptocurrency. This includes fund tracking, fiat to cryptocurrency fund conversion, transaction logging, account creation, voting scenarios, bootstrapping mode, role assignment, and the ability of accounts to take special actions given their roles (e.g., law enforcement account freezing and central bank fund creation). Certain aspects are simplified, such as monetary policy options, as our goal was not to create a production system but to demonstrate that this managed architecture approach is feasible. We tested our code by deploying it to a local Ethereum test environment.

A couple of money creation schemes are provided in our system as examples. Schemes are provided to vote and carry out money creation in arbitrary accounts (following a top-down approach) as well as in all user accounts in the form of interests (following a bottom-up approach). Furthermore, these schemes can be either push-based or pull-based. In the push-based model, the money creation function creates funds in the recipient(s) account without any action being required from the recipient(s). In the pull-based model, the currency managers set rules through a single transaction to give the right to the recipient(s) to create their own funds according to this set of rules. This can provide scalability gains as users do not have to claim their allowance right away, and instead, may wait until they need it without any risk of not receiving it. In the case of periodic funds creation (e.g. interests, dividends), a user might be able to skip claiming funds between period X and period X + Y, and then, withdraw funds at period X + Y + 1 for all of the periods between X and X + Y + 1 combined. Finally, a set of view functions allows one to selectively control the visibility of monetary creation and other on-chain fund data, both at the user level and at the currency management level (e.g., global supply indicators).

Note that our implementation did not cover the off-chain aspects of our cryptocurrency architecture. In particular, we did not build the security or visibility gateways (although we did write the smart contract view functions to support the latter). We also did not modify the Ethereum mining software to only publish blocks in collaboration with the validators specified by the smart contracts (but we did implement the smart contract code to enable a set of validators to use the on-chain data to manage themselves).

VIII. Related Work

Most of the existing work related to managed cryptocurrencies consists in studies and pilots on blockchain-based central bank digital currencies (CBDC), as well as research and development of protocols for stablecoins, algorithmic currency management, and privacy-preserving KYC/AML checks.

The literature distinguishes two main categories of CBDCs: wholesale and retail. As explained by the Bank for International Settlements (BIS) in their money taxonomy [16], a wholesale CBDC is only available to financial institutions and mainly intended for inter-bank transactions whereas a retail CBDC is globally accessible and usable by the general public. Our managed cryptocurrency architecture is geared towards retail CBDCs, although it could also be launched as a wholesale CBDC, at least initially.

We have developed our architecture to change how cryptocurrencies usually work to enable support for retail CBDCs. As stated earlier, this does not mean that we are necessarily claiming that one should be created; we are simply providing some of the technical capability to do so. That said, the subject of state or central bank issued cryptocurrencies has been one of considerable interest. A BIS poll in 2018 showed that more than 70 percent of central banks worldwide were already engaged in CBDC work [17]. Some central banks, such as the central bank of Canada (project Jasper [18]), the Monetary Authority of Singapore (project Ubin [19]), or the Bank of Thailand [20] have focused their research on wholesale CBDC. Also, the European Central Bank and the Bank of Japan are conducting joint research on wholesale CBDCs with Project Stella [21]. Others, such as the Ecuadorian Central Bank ('Dinero Electronico' [22]), the People's Bank of China [23], and the Government of Venezuela (Petro [24]) aim at developing a CBDC for retail use. It should be noted, however, that many central bank efforts do not use (nor plan to use in the future) distributed ledger technologies (DLT); for example the Sveriges Riksbank from Sweden [25] stated that they currently deemed DLT too inefficient for use in a retail CBDC. An example of a non-DLT retail electronic currency is the e-Peso [26]. This is a pilot from the Central Bank of Uruguay that was launched as complement to physical cash but relied on a central registry for ownership recording.

Aside from efforts from governments and central banks, several other blockchain-based research projects have entered the field of managed cryptocurrencies. For example, RScoin [27] provides a cryptocurrency framework using a UTXO model where generation of the monetary supply is controlled by a central authority and transaction processing is handled by dedicated institutions, called 'mintettes'; ultimately, the central authority handles the creation and posting of new blocks. Our system differs from this approach in that currency managers do not influence the block creation process nor benefit from special viewing rights over the content of the blockchain. Another example of a managed cryptocurrency system can be found in Fedcoin [28], which builds upon RScoin's framework by providing a Node.js implementation, KYC rules that enable a central bank to blacklist users, and improved anonymity features. However, unlike our proposal, it does not natively offer the ability for accounts to be assigned roles; this leaves the central bank as the sole entity involved in the identity provider, management, and system security functions (e.g., identity-proofing new accounts, freezing unlawful users, and coin production).

'Decentralized Finance' projects, many of which are currently built with smart contracts deployed on the public Ethereum blockchain or as second layer solutions atop Bitcoin, are also being developed for stablecoins and decentralized currency management where money supply is governed algorithmically (such as Dai [29]). In our system, unlike reserve-backed stablecoins (such as Libra [30], USD Coin [31], and J.P. Morgan Coin [32]) that are pegged one-to-one with the asset(s) that they represent, there is a built-in currency manager role that can develop monetary instruments and vote for monetary policies to increase and decrease the currency supply. Since it is programmable, novel, potentially more flexible monetary instruments may be implemented.

From a security point of view, efforts are being made to offer security standards, toolsets, and services for cryptocurrencies. For example, EIP-1080 [33] is an Ethereum standard that offers an interface geared towards charge back and theft prevention/resolution for ERC-20 tokens [14]. Also, more loosely related is that the Enterprise Ethereum Alliance (EEA) Legal Industry Working Group [34] intends to standardize law-compliant smart contract designs.

Our system provides user privacy through use of a permissioned blockchain that supports roles with different responsibilities and data visibility (e.g., block publishers cannot see account owners' identities). However, cash transactions offer an ideal for anonymity and attempts to achieve this ideal for electronic currencies have been the subject of much research. The development of some privacy-preserving technologies, such as zero-knowledge protocols, has assisted in this objective. For example, Chaum introduced eCash [35] in 1983, one of the first attempts at anonymizing electronic money transactions via the use of blind signatures; Zcash [36] is an example of cryptocurrency that relies on a type of zero-knowledge proof called zk-SNARKs for keeping transactions private; and ChainAnchor [37] offers a method based on the 'Enhanced Privacy ID' zero-knowledge protocol for controlling access to a permissioned blockchain while allowing users to transact pseudonymously and maintain transaction unlinkability.

IX. Conclusion

Most cryptocurrencies and cryptocurrency research efforts focus on providing cryptocurrencies with strong anonymity and privacy guarantees in a robust distributed system that is not owned or managed by any single entity or group. We do not dispute the importance of such efforts and the emergence of the associated new social constructs, but point out that research in managed cryptocurrencies integrated with our current institutions has been sorely lacking. This is unfortunate as all people live under the laws of their respective countries and it is thus important to research cryptocurrencies that can explicitly support those laws.

In recent years, central banks have been interested in this area, but some of their researchers have discounted cryptocurrency solutions because the foundational technology appears incompatible with central bank goals, especially the support for KYC and AML laws. In this work, we showed how the foundational elements of a cryptocurrency can be rethought to support central bank goals and to explicitly support the laws that apply to electronic fiat currencies. We hope to convince the reader that this type of approach is technically feasible and that cryptocurrencies can be developed that integrate with an associated fiat currency and explicitly support the laws of the respective government.

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[2] A. Berentsen and F. Schar, "The case for central bank electronic money and the non-case for central bank cryptocurrencies," 2018. [Online]. Available: https://doi.org/10.20955/r.2018.97-106

[3] G. Greenspan, "Multichain private blockchain," *White paper*, 2015. [Online]. Available: https://www.multichain.com/download/MultiChain-White-Paper.pdf

[4] J. Barrdear and M. Kumhof, "The macroeconomics of central bank issued digital currencies," 2016. [Online]. Available: https://www.bankofengland.co.uk/working-paper/2016/the-macroeconomics-of-central-bank-issued-digital-currencies

[5] P. Mell, "Managed blockchain based cryptocurrencies with consensus enforced rules and transparency," in *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 2018, pp. 1287–1296.

[6] Parity, "Parity tech documentation - validator sets webpage," Retrieved on 2 July 2019 from https://wiki.parity.io/Validator-Set.htmlreporting-contract.

[7] P. Network, "Proof of authority: consensus model with identity at stake," 2017. [Online]. Available: https://medium.com/poa-network/proof-of-authority-consensus-model-with-identity-at-stake-d5bd15463256

[8] Microsoft, "Ethereum proof-of-authority consortium," 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/blockchain/templates/ethereum-poa-deployment

[9] Anonymous, "Governing decentralization: How on-chain voting protocols operate and vary," 2018. [Online]. Available: https://cointelegraph.com/news/governing-decentralization-how-on-chain-voting-protocols-operate-and-vary

[10] E. EIP-1202, "Voting standard," 2018. [Online]. Available: https://github.com/ethereum/EIPs/blob/master/EIPS/eip-1202.md

[11] Aragon, "Aragon website," Retrieved on 10 June 2019 from https://aragon.org/.

[12] Bitshares, "Bitshares website," Retrieved on 3 July 2019 from https://bitshares.org/.

[13] M. Vukolić, "Rethinking permissioned blockchains," in *Proceedings of the ACM Workshop on Blockchain, Cryptocurrencies and Contracts*. ACM, 2017, pp. 3–7.

[14] F. Vogelsteller and V. Buterin, "Erc-20 token standard," 2015. [Online]. Available: https://eips.ethereum.org/EIPS/eip-20

[15] L. Lesavre, P. Varin, P. Mell, M. Davidson, and J. Shook, "A taxonomic approach to understanding emerging blockchain identity management systems (draft)," National Institute of Standards and Technology, Tech. Rep., 2019.

[16] M. L. Bech and R. Garratt, "Central bank cryptocurrencies," *BIS Quarterly Review September*, 2017.

[17] C. Barontini and H. Holden, "Proceeding with caution-a survey on central bank digital currency," *Proceeding with Caution-A Survey on Central Bank Digital Currency (January 8, 2019). BIS Paper*, no. 101, 2019.

[18] B. of Canada, "Fintech experiments and projects," Retrieved on 26 June 2019 from https://www.bankofcanada.ca/research/digital-currencies-and-fintech/fintech-experiments-and-projects/.

[19] M. A. of Singapore, "Project ubin: Central bank digital money using distributed ledger technology," Retrieved on 25 June 2019 from https://www.mas.gov.sg/schemes-and-initiatives/Project-Ubin.

[20] B. of Thailand, "Public vs private blockchain in a nutshell," 2019. [Online]. Available: https://www.bot.or.th/Thai/PressandSpeeches/Press/News2562/n562e.pdf

[21] E. C. B. . B. O. Japan, "Boj/ecb joint research project on distributed ledger technology," 2018. [Online]. Available: https://www.boj.or.jp/en/announcements/release_2019/rel190604a.htm

[22] J. Campuzano, G. Cruz, and G. Jr. Y Maza Iñiguez, "El fracaso del dinero electrónico en ecuador," vol. 7, pp. 82–101, 08 2018.

[23] W. Knight, "Mit technology review - china's central bank has begun cautiously testing a digital currency," 2017. [Online]. Available: https://www.technologyreview.com/s/608088/chinas-central-bank-has-begun-cautiously-testing-a-digital-currency

[24] G. of Venezuela, "Petro webpage," Retrieved on 28 June 2019 from https://www.petro.gob.ve/eng/home.html.

[25] S. Riksbank, "E-krona webpage," Retrieved on 26 June 2019 from https://www.riksbank.se/en-gb/payments–cash/e-krona/, 2019.

[26] I. M. Fund, "Uruguay : 2018 article iv consultation-press release; staff report; and statement by the executive director for republic of uruguay," 2019. [Online]. Available: https://bit.ly/2SkS7pE

[27] G. Danezis and S. Meiklejohn, "Centrally banked cryptocurrencies," *arXiv preprint arXiv:1505.06895*, 2015.

[28] S. Gupta, P. Lauppe, and S. Ravishankar, "Fedcoin: A blockchain-backed central bank cryptocurrency," 2017. [Online]. Available: https://zoo.cs.yale.edu/classes/cs490/16-17b/gupta.sahil.sg687

[29] Maker, "Dai website," Retrieved on 18 June 2019 from https://makerdao.com/en/dai/.

[30] Libra, "Libra white paper," 2019. [Online]. Available: https://libra.org/en-US/wp-content/uploads/sites/23/2019/06/LibraWhitePaper_en_US.pdf

[31] Coinbase, "Usdc webpage," Retrieved on 18 June 2019 from https://www.coinbase.com/usdc.

[32] J. Morgan, "J.p. morgan creates digital coin for payments," 2019. [Online]. Available: https://www.jpmorgan.com/global/news/digital-coin-payments

[33] E. EIP-1080, "Recoverable token," 2018. [Online]. Available: https://github.com/ethereum/EIPs/blob/master/EIPS/eip-1080.md

[34] E. E. Alliance, "Eea legal industry working group press release," 2017. [Online]. Available: https://entethalliance.org/ethereum-enterprise-alliance-legal-industry-working-group-press-release-2/

[35] D. Chaum, "Blind signatures for untraceable payments," in *Advances in cryptology*. Springer, 1983, pp. 199–203.

[36] E. C. Company, "Zcash webpage," Retrieved on 20 June 2019 from https://z.cash/.

[37] T. Hardjono, N. Smith, and A. S. Pentland, "Anonymous identities for permissioned blockchains," 2014. [Online]. Available: https://petertodd.org/assets/2016-04-21/MIT-ChainAnchor-DRAFT.pdf

Estimating the Parameters of Circles and Ellipses Using Orthogonal Distance Regression and Bayesian Errors-in-Variables Regression

Jolene Splett* Amanda Koepke* Felix Jimenez†

Abstract
In ordinary least-squares regression, independent variables are assumed to be known without error. However, in many real-life situations this assumption is not valid. Both orthogonal distance regression and Bayesian errors-in-variables regression can be used to estimate model parameters when there are errors in the dependent and independent variables.

To illustrate the use of the maximum-likelihood and Bayesian approaches, we use both methods to estimate the parameters of a circle. The data used for circle fitting were taken from the cross section of an optical fiber. The shape of optical fibers is important when joining two fibers, so accurate dimensional measurements are critical to minimizing coupling loss. We then compare the results of the two techniques when fitting an ellipse to simulated data.

Circle and ellipse fitting using maximum-likelihood methods have been well documented; however, Bayesian methods for these tasks are less developed. As expected, we found that the Bayesian approach for circle fitting is more intuitive and easier to implement than the maximum-likelihood approach, but generalizing the Bayesian approach to ellipse fitting was surprisingly difficult.

Key Words: Bayesian statistics, errors-in-variables regression, orthogonal distance regression, circle fitting, ellipse fitting

1. Introduction

In many real-life regression problems, errors are present in both the independent and dependent variables. Circles and ellipses are classic examples of data that have errors in both the x and y variables.

We examine two methods of fitting circle and ellipse data. The first method is orthogonal distance regression (ODR). In ODR, maximum-likelihood estimates are obtained by minimizing distances between data points and the fitted curve. The second method is Bayesian errors-in-variables regression (EIV). In general, the Bayesian technique allows for easy and intuitive explicit modeling of errors in both the x and y variables for simple models. In this paper, we seek to extend the Bayesian EIV method for the purpose of estimating the parameters of circles and ellipses. While maximum-likelihood based approaches to the circle-fitting problem are well documented (Boggs et al. [1992], Chernov [2011]), very little has been written about a Bayesian approach to this classic errors-in-variables problem.

Data representing the cross section of an optical fiber (Wang et al. [1997], Mamileti et al. [1992]) are used to demonstrate the ODR and EIV methods for estimating the parameters of a circle. We then use both methods to fit an ellipse to simulated data.

Sections 2 and 3 describe the data being fit and detail our circle fitting and ellipse fitting techniques. Results are provided in Section 4, and Section 5 summarizes our findings.

*National Institute of Standards and Technology, 325 Broadway MC898.03, Boulder, CO 80305
†University of Colorado, Department of Physics, 390 UCB, Boulder, CO 80309

Splett, Jolene D.; Jimenez, Felix; Koepke, Amanda. "Estimating the Parameters of Circles and Ellipses Using Orthogonal Distance Regression and Bayesian Errors-in-Variables Regression." Paper presented at 2019 Joint Statistical Meetings, Denver, CO, US. July 28, 2019 - August 01, 2019.

2. Circle Fitting

2.1 Optical Fiber Data

Dimensional measurements of an optical fiber cross section are obtained by circle fitting. Accurate dimensional measurements are critical to providing the best possible transmission of light through the fiber; as much as 1 μm offset in joining two fibers will result in about 5 % loss of signal (Wang et al. [1997]).

The data were obtained by first generating a gray-scale image of a fiber cross section (Figure 1). Next, an edge-detection algorithm was applied to the image to define the outside edge of the fiber, producing the (x, y) coordinate pairs required for the analysis.

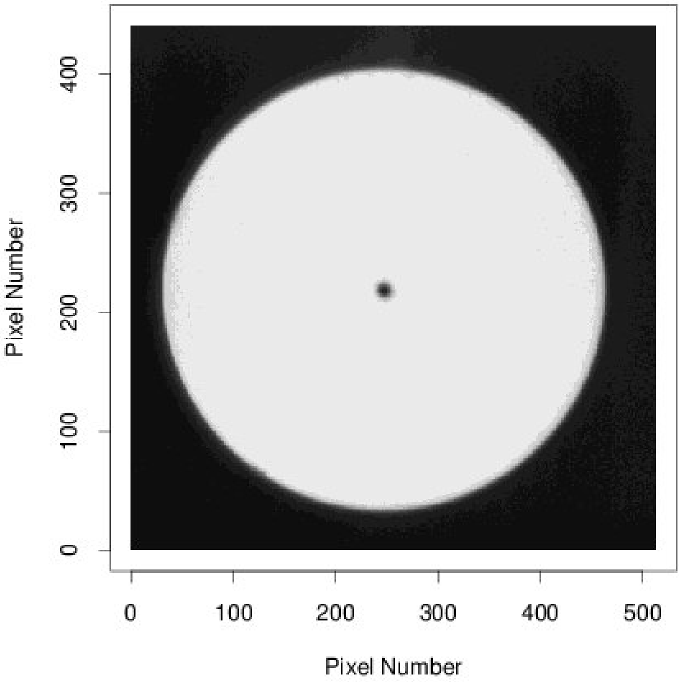


Figure 1: Gray-scale image of an optical-fiber cross section. Data and figure from Wang et al. [1997].

For reasons discussed in 2.2.1, two different representations of a circle are used to perform model fitting. For ODR, we use the general equation of a circle, $r^2 = (x_i - x_c)^2 + (y_i - y_c)^2$. The parameters of interest are the radius (r), the x-coordinate of the center (x_c), and the y-coordinate of the center (y_c).

For EIV, we define a general circle in terms of (x, y) coordinates. Consider a unit circle, centered at $(0, 0)$ with radius $r = 1$. We denote points on the unit circle as $(\tilde{x}_i, \tilde{y}_i)$, where $\tilde{x}_i = \cos(\theta_i)$, $\tilde{y}_i = \sin(\theta_i)$, and $\theta_i \in (-\pi, \pi]$. From the unit circle, we can obtain points on a standard circle (x'_i, y'_i) by multiplying by a constant r, so $x'_i = r\cos(\theta_i)$ and $y'_i = r\sin(\theta_i)$. This standard circle is centered at zero and has radius r. We can move this circle to a new center (x_c, y_c), which gives us the general form for points on a circle

$$(X_i = x_c + r\cos(\theta_i),\ \ Y_i = y_c + r\sin(\theta_i))\,. \tag{1}$$

2.2 Bayesian Estimates

2.2.1 *The model*

To obtain parameter estimates in the frequentist framework, the sum-of-squared residuals is minimized subject to the constraint from the general equation of a circle; creating a similar framework for Bayesian EIV regression is not obvious. Keksel et al. [2018] modeled a circle using Bayesian methods by defining their likelihood using $(x_i - x_c)^2 + (y_i - y_c)^2 - r^2$, assigning this a normal distribution with zero mean and some variance, but the approach does not allow the errors in the x and y variables to be modeled separately. The method used by Werman and Keren [2001] also has this limitation. Since the ODR circle parameterization makes it difficult to determine the likelihood for the Bayesian EIV regression approach, we propose a more intuitive parameterization, rewriting the model in terms of x and y coordinates.

We use the general form of a circle (1) to model our observed data. First, for notational and computational convenience, we roughly center our observed circle by subtracting the means, so our new data are represented as $x_{i,new} = x_i - \bar{x}$ and $y_{i,new} = y_i - \bar{y}$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. The center of the new circle is then $x_0 = x_c - \bar{x}$ and $y_0 = y_c - \bar{y}$, where (x_c, y_c) is the center of the observed circle data as defined in Section 2.1. Thus for the EIV model, (1) becomes

$$(X_i^* = x_0 + r\cos(\theta_i), \;\; Y_i^* = y_0 + r\sin(\theta_i))\,. \tag{2}$$

We assume $x_{i,new}$ and $y_{i,new}$ are normally distributed about the points of some true circle defined by (2) with variances σ_x^2 and σ_y^2, so

$$x_{i,new} \sim N(X_i^*, \sigma_x^2) \qquad y_{i,new} \sim N(Y_i^*, \sigma_y^2). \tag{3}$$

Now we must assign prior distributions to all of the parameters in our model. We know that the deviations of the points about the circle are small, encouraging us to use relatively concentrated prior distributions for σ_x and σ_y. We assume tight gamma prior distributions for σ_x and σ_y,

$$\sigma_x \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 50) \qquad \sigma_y \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 50).$$

Since we center our data, x_0 and y_0 should be close to zero, so we assume *a priori*

$$x_0 \sim N(0, 1) \qquad y_0 \sim N(0, 1).$$

The radius is required to be positive, and we assume *a priori*

$$r \sim N(60, 10^2) I_{r \geq 0}.$$

This prior distribution is centered about a realistic value of r, but has large variance.

We use a von Mises prior distribution for the θ_i, bounded between $-\pi$ and π, with location equal to zero and concentration, κ, equal to 0.1. This distribution is plotted for various concentration values in Figure 2. The concentration determines how concentrated the distribution is about the location. The distribution becomes uniform over the interval as κ goes to zero. However, in practice the concentration is restricted to be greater than zero for computational reasons, so we give κ a small value. The von Mises distribution is a common circular distribution that has desirable inferential properties (Mardia and Jupp [2000]).

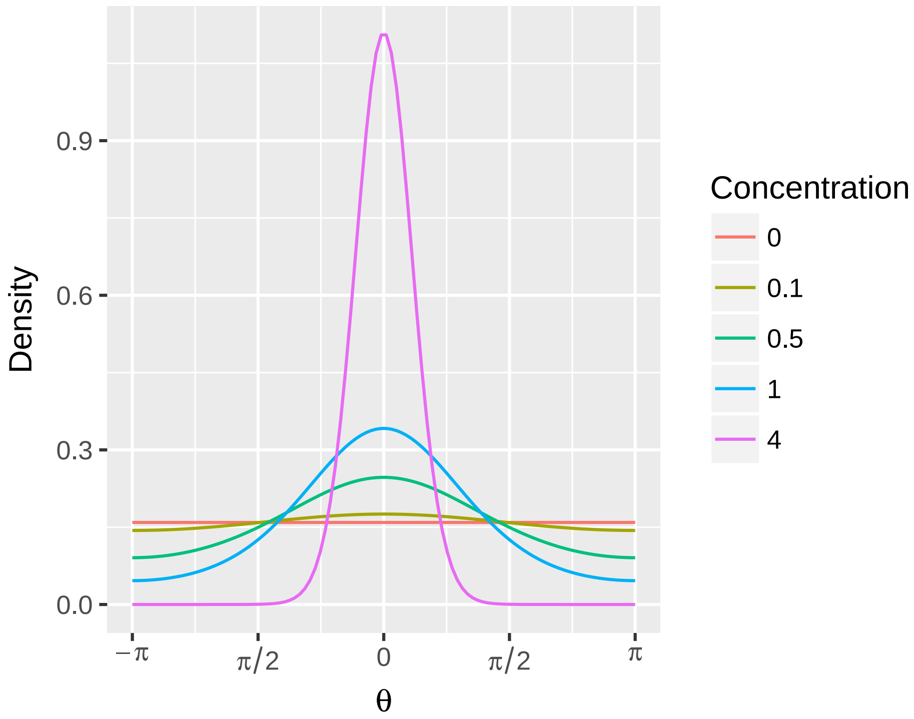


Figure 2: Probability density functions for von Mises distributions on θ, bounded between $-\pi$ and π, with different values for the concentration. As the concentration approaches zero, the distribution becomes more uniform.

2.2.2 MCMC

Now that a model, a set of priors, and the observed data have been defined, we turn our attention to estimating the parameters. To accomplish this, we use Hamiltonian Monte Carlo (HMC), implemented via `Stan` using the `R` package `rstan` [Carpenter et al., 2017, Stan Development Team, 2018], to sample from the posterior distribution of the parameters given the observed data. In particular, we run three chains, each with a burn in of 500 samples and a total of 1000 iterations.

The large number of parameters make this model difficult to fit without reasonable initial values for HMC. We initialize all three chains as follows. The initial radius value is $r_{init} = 60$, the standard deviations (σ_x, σ_y) are both initialized at 0.01, and the center is $(x_{0,init}, y_{0,init}) = (0, 0)$. To initialize the θ_i, we relate them to the data $(x_{i,new}, y_{i,new})$ and the other initial values, so

$$\theta_{i,init} = \arctan\left(\frac{\tilde{y_i}}{\tilde{x_i}}\right) = \arctan\left(\frac{(y_{i,new} - y_{0,init})/r_{init}}{(x_{i,new} - x_{0,init})/r_{init}}\right),$$

which reduces to $\theta_{i,init} = \arctan\left(\frac{y_{i,new}}{x_{i,new}}\right)$.

2.3 Maximum-Likelihood Estimates

We obtain maximum-likelihood estimates of the parameters for the circle model using the Fortran package, ODRPACK (Boggs et al. [1992]). The model for circle fitting,

$$0 = (x_i - x_c)^2 + (y_i - y_c)^2 - r^2, \tag{4}$$

is an implicit model because there is no explicit independent variable. For implicit multivariate orthogonal distance regression using ODRPACK, define $z_i = (x_i, y_i)$ and $\delta_i = (\delta_{x_i}, \delta_{y_i})$, where δ_{x_i} and δ_{y_i} represent the errors in x_i and y_i, respectively. Parameters are estimated using

$$\min_{\beta,\delta} \sum_{i=1}^{n} w_{\delta_i} {\delta_i}^2 \tag{5}$$

subject to the constraint

$$f_i(z_i + \delta_i; \beta) = ((x_i - \delta_{x_i}) - x_c)^2 + ((y_i - \delta_{y_i}) - y_c)^2 - r^2 = 0, \qquad i = 1, \ldots, n,$$

where w_{δ_i} are weights and $\beta = (x_c,\, y_c,\, r)$. For our problem, $w_{\delta_i} = 1$ for all i.

3. Ellipse Fitting

3.1 Simulated Data

To understand the models for fitting the ellipse, we review some geometry. An ellipse can be defined in terms of two circles with radii r_x and r_y, with $r_x > r_y$. This is depicted in Figure 3. As with the circle model, we first consider the unit circle (green circle in the figure). In the figure, A denotes a point on the unit circle, and $A = (\tilde{x_i}, \tilde{y_i}) = (\text{cos}(\theta_i), \text{sin}(\theta_i))$, where $\theta_i \in [0, 2\pi)$. Using points on the unit circle, we obtain points on the red circle by multiplying by r_y, so point B in the figure equals $(r_y\text{cos}(\theta_i), r_y\text{sin}(\theta_i))$. Similarly, we obtain points on the blue circle by multiplying by r_x, so point C in the figure equals $(r_x\text{cos}(\theta_i), r_x\text{sin}(\theta_i))$. Using these two circles, we obtain a point (D) on the standard ellipse as $(x_i', y_i') = (r_x\text{cos}(\theta_i), r_y\text{sin}(\theta_i))$. The standard ellipse is centered at zero, has major axis r_x, and has minor axis r_y.

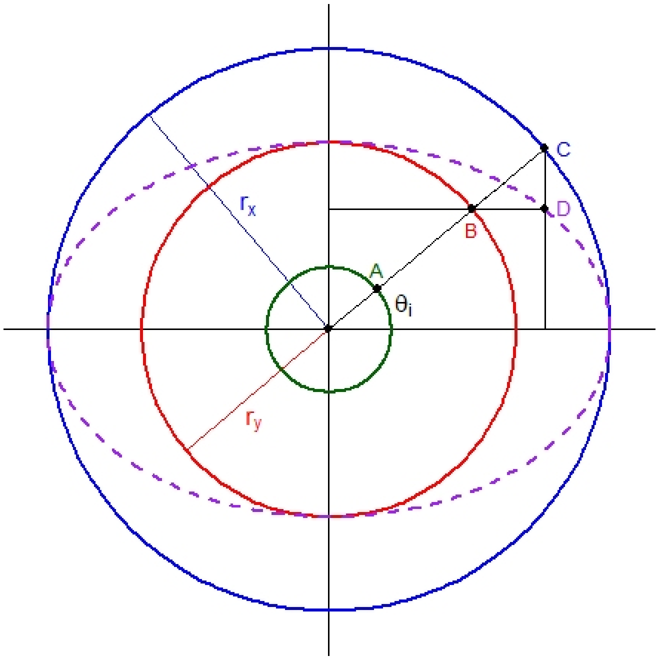


Figure 3: Geometric interpretation of an ellipse based on two circles with different radii.

The standard ellipse can be rotated by angle α and moved to a new center (x_c, y_c), which gives us the general form for points on an ellipse

$$\begin{aligned} X_i &= x_c + r_x \cdot \cos(\theta_i) \cdot \cos(\alpha) - r_y \cdot \sin(\theta_i) \cdot \sin(\alpha) \\ Y_i &= y_c + r_y \cdot \sin(\theta_i) \cdot \cos(\alpha) + r_x \cdot \cos(\theta_i) \cdot \sin(\alpha). \end{aligned} \tag{6}$$

For ellipse fitting, we simulate 1000 noisy data pairs as $x_i \sim N(X_i, \sigma_x^2)$ and $y_i \sim N(Y_i, \sigma_y^2)$ using the values shown in Table 1. The $\theta_1, \ldots, \theta_{1000}$ are an evenly spaced sequence from $-\pi + \epsilon$ to $\pi - \epsilon$, with $\epsilon = 0.00001$. Figure 4 displays the data points used in the analysis.

Table 1: Parameters used to generate simulated ellipse data.

Parameter	Value
x_c	71
y_c	74
r_x	62
r_y	50
α	-0.55 rad
σ_x	0.05
σ_y	0.07

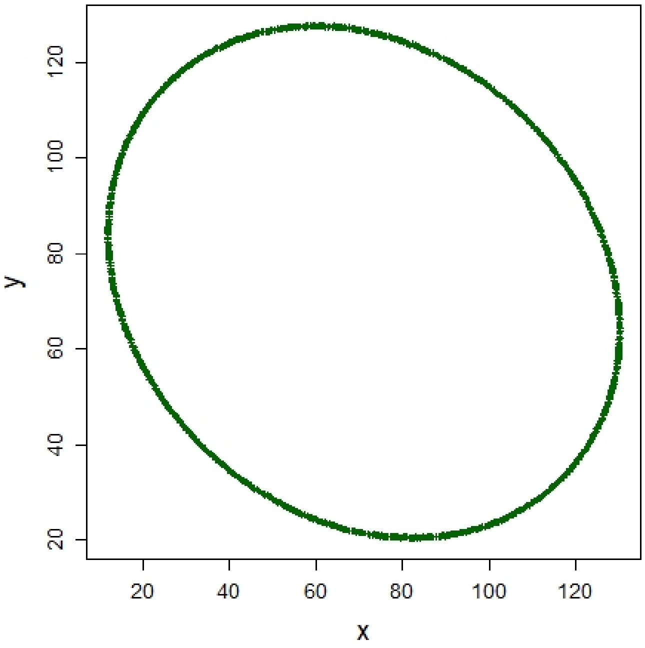


Figure 4: Simulated ellipse data (n=1000) generated using (6) and the parameters listed in Table 1.

3.2 Bayesian Estimates

3.2.1 The Model

As before, we use the general form of the ellipse (6) to model our observed data. We again roughly center our observed ellipse by subtracting the means, so our new data are represented as $x_{i,new} = x_i - \bar{x}$ and $y_{i,new} = y_i - \bar{y}$, and the center of the new ellipse is $x_0 = x_c - \bar{x}$ and $y_0 = y_c - \bar{y}$, where (x_c, y_c) is the center of the observed ellipse data as defined in Section 3.1. Thus for the EIV model, (6) becomes

$$\begin{aligned} X_i^* &= x_0 + r_x \cdot \cos(\theta_i) \cdot \cos(\alpha) - r_y \cdot \sin(\theta_i) \cdot \sin(\alpha) \\ Y_i^* &= y_0 + r_y \cdot \sin(\theta_i) \cdot \cos(\alpha) + r_x \cdot \cos(\theta_i) \cdot \sin(\alpha). \end{aligned} \tag{7}$$

As with the circle model, we assume $x_{i,new}$ and $y_{i,new}$ are normally distributed about the points of some true ellipse defined by X_i^* and Y_i^*, so

$$x_{i,new} \sim N(X_i^*, \sigma_x^2) \qquad y_{i,new} \sim N(Y_i^*, \sigma_y^2). \tag{8}$$

We use many of the same prior distributions assumed in Section 2.2. Since we center our data, we assume *a priori* that the center of our ellipse should be close to (0,0), so

$$x_0 \sim N(0,1) \qquad y_0 \sim N(0,1).$$

The standard deviations of x_i and y_i, both of which are positive, are defined by

$$\sigma_x \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 50) \qquad \sigma_y \sim \text{Gamma}(\text{shape} = 2, \text{rate} = 50).$$

Both the x radius (r_x) and y radius (r_y) must be positive. We assume

$$r_x \sim N(60, 10^2) I_{[r_x>0]} \qquad r_y \sim N(60, 10^2) I_{[r_y>0]}.$$

Again, this prior distribution expresses a lot of uncertainty about these parameter values, even though we think they should be around 60. The larger of the r_x and r_x values corresponds to the major axis (M) and the smaller value corresponds to the minor axis (m).

The rotation of the ellipse, α, is constrained to cover a 90 degree range to ensure a unique solution for the parameter estimates. This restriction on α allows us to avoid imposing a restriction on the relative sizes of r_x and r_y due to the fact that the major axis must be greater than the minor axis. If α covered an entire 180 degree range, then two different values of α could be used to describe the same ellipse if the major and minor axes are switched. *A priori* we assume α has a uniform distribution between $-\pi/2$ and zero.

Finally, we again use a von Mises prior distribution for the θ_i, bounded between $-\pi$ and π, with location equal to zero and concentration equal to 0.1.

3.2.2 MCMC

Our results are very sensitive to initial values, particularly for the angle of rotation. Thus, we use the data to estimate a rough angle of rotation to initialize the model. Using the roughly centered data $(x_{i,new}, y_{i,new})$, we calculate $r_i = \sqrt{x_{i,new}^2 + y_{i,new}^2}$. The maximum r_i should correspond to a point close to the major axis, so the angle that corresponds to that point, constrained to be between $-\pi$ and zero, is a good initial value for the angle of rotation of our ellipse. Specifically, $\alpha_{init} = \text{arctan}\left(\frac{y_{j,new}}{x_{j,new}}\right)$, where j is the index for the maximum value of the r_i. If this is not between $-\pi/2$ and zero, we find the corresponding angle that is in this range.

For the center, we use $(x_{0,init}, y_{0,init}) = (0, 0)$. Given $(x_{i,new}, y_{i,new})$, we obtain a standard ellipse roughly centered at $(0, 0)$ and rotated by angle, $-\alpha_{init}$, using

$$\begin{aligned} x_i' &= (x_{i,new} - x_{0,init}) \cdot \text{cos}(-\alpha_{init}) - (y_{i,new} - y_{0,init}) \cdot \text{sin}(-\alpha_{init}) \\ y_i' &= (y_{i,new} - y_{0,init}) \cdot \text{cos}(-\alpha_{init}) + (x_{i,new} - x_{0,init}) \cdot \text{sin}(-\alpha_{init}). \end{aligned} \tag{9}$$

From the standard ellipse (9), we can initialize r_x and r_y as

$$\begin{aligned} r_{x,init} &= \frac{|\text{min}(x_i')| + |\text{max}(x_i')|}{2} \\ r_{y,init} &= \frac{|\text{min}(y_i')| + |\text{max}(y_i')|}{2}. \end{aligned}$$

To initialize the θ_i, we set

$$\theta_{i,init} = \text{arctan}\left(\frac{y_i'/r_{y,init}}{x_i'/r_{x,init}}\right).$$

The standard deviations σ_x and σ_y are both initialized at 0.01. We use these values to initialize three chains, each with a burn in of 300 and a total of 600 iterations, again using HMC to sample from the posterior distribution.

3.3 Maximum-Likelihood Estimates

A general formula for an ellipse that is translated and rotated in the x, y plane (Fuller [1987]) is

$$\beta_1(y_i - y_c)^2 + 2\beta_2(y_i - y_c)(x_i - x_c) + \beta_3(x_i - x_c)^2 - 1 = 0. \tag{10}$$

The center of the ellipse is $(x_c,\, y_c)$; however, the remaining parameters, $(\beta_1,\, \beta_2,\, \beta_3)$, have no geometric interpretation. Thus, we compute the angle of rotation, the major axis, and minor axis, $(\alpha,\, M,\, m)$, from the estimated model parameters $(\hat{x}_c,\, \hat{y}_c,\, \hat{\beta}_1,\, \hat{\beta}_2,\, \hat{\beta}_3)$.

We use ODRPACK to obtain the maximum-likelihood estimates of the ellipse model parameters by the same procedure described in Section 2.3. As in circle fitting for an implicit model (10), ODRPACK minimizes (5) subject to

$$\begin{aligned} g_i(z_i + \delta_i; \beta) &= \beta_1((y_i - \delta_{y_i}) - y_c)^2 + 2\beta_2((y_i - \delta_{y_i}) - y_c)((x_i - \delta_{x_i}) - x_c) \\ &+ \beta_3((x_i - \delta_{x_i}) - x_c)^2 - 1 = 0, \;\; i = 1, \ldots, n, \end{aligned}$$

where $z_i = (x_i,\, y_i)$ and $\beta = (x_c,\, y_c,\, \beta_1,\, \beta_2,\, \beta_3)$.

The standard errors of $(\hat{x}_c,\, \hat{y}_c)$ are obtained directly from ODRPACK. However, the standard errors of the additional parameters of interest, $(\hat{\alpha},\, \hat{M},\, \hat{m})$, are computed from Monte Carlo simulations (Lafarge and Possolo [2015]) based on $\hat{\beta}$ and the estimated variance-covariance matrix of $\hat{\beta}$.

4. Results

Parameter and interval estimates determined by maximum-likelihood and Bayesian methods are shown in Tables 2 and 3 for the circle and ellipse models, respectively. There is very little difference between estimated parameters for maximum-likelihood and Bayesian methods for both circle and ellipse models. Plots of the parameter estimates and their 95 % confidence intervals and 95 % credible intervals are shown in Figures 5 and 6.

Table 2: Circle parameters estimated by ODR and Bayesian EIV.

Parameter	ODR Estimate (95 % Confidence Interval)	Bayesian EIV Estimate (95 % Credible Interval)
x_0, μm	71.89 (71.89, 71.90)	71.90 (71.89, 71.90)
y_0, μm	74.26 (74.26, 74.27)	74.27 (74.26, 74.27)
r, μm	62.47 (62.47, 62.47)	62.48 (62.47, 62.48)

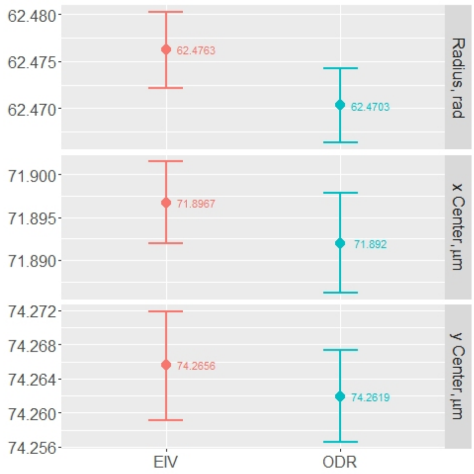
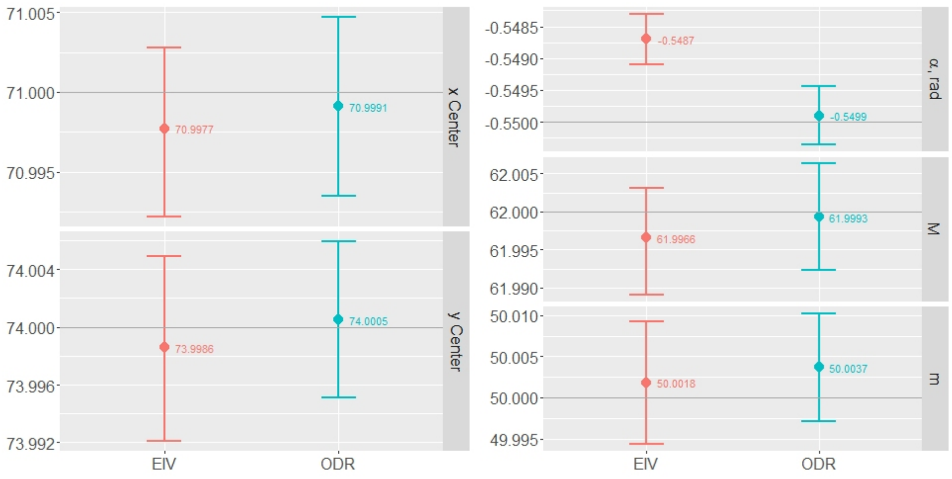


Figure 5: Bayesian and maximum-likelihood interval estimates for circle parameters.

Table 3: Ellipse parameters estimated by ODR and Bayesian EIV.

Parameter	ODR Estimate (95 % Confidence Interval)	Bayesian EIV Estimate (95 % Credible Interval)
x_0	70.9991 (70.9935, 71.0047)	70.9977 (70.9922, 71.0028)
y_0	74.0005 (73.9952, 74.0058)	73.9986 (73.9921, 74.0049)
α, rad	-0.5499 (-0.5505, -0.5496)	-0.5487 (-0.5491, -0.5483)
M	61.9993 (61.9924, 62.0062)	61.9966 (61.9891, 62.0031)
m	50.0037 (49.9973, 50.0102)	50.0018 (49.9943, 50.0093)



(a)

(b)

Figure 6: Bayesian and maximum-likelihood interval estimates for ellipse parameters. The horizontal gray reference lines indicate the true parameter values.

Splett, Jolene D.; Jimenez, Felix; Koepke, Amanda. "Estimating the Parameters of Circles and Ellipses Using Orthogonal Distance Regression and Bayesian Errors-in-Variables Regression." Paper presented at 2019 Joint Statistical Meetings, Denver, CO, US. July 28, 2019 - August 01, 2019.

Figure 7 displays a portion of the measurement data and the circle fitting results for both methods. Plotted on the scale of the raw data, the two methods produce nearly identical circles. The plot shows how the raw data deviates from a circle due to imperfections in the optical fiber.

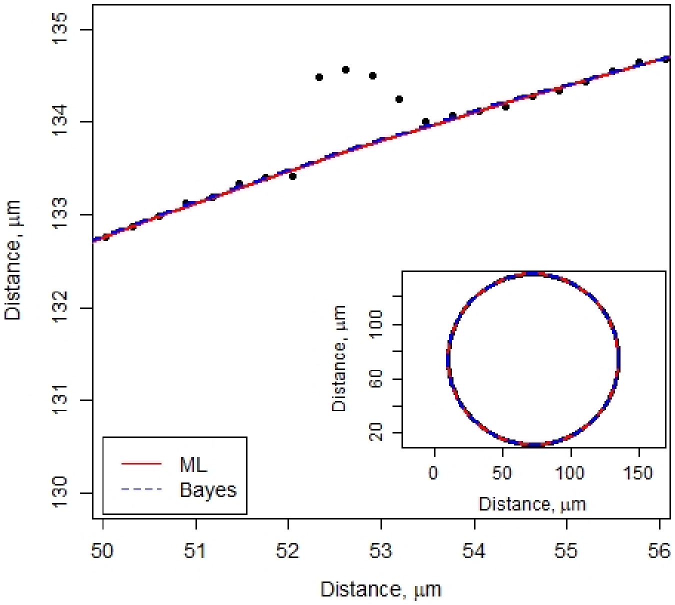


Figure 7: Maximum likelihood and Bayesian circle fitting results.

5. Discussion

We fit a circle model to a real data set and an ellipse model to a simulated data set using both maximum-likelihood and Bayesian approaches. The two methods produced similar parameter and interval estimates. The Bayesian approach is more intuitive for the error-in-variables problem than the maximum-likelihood method. Although we expected the Bayesian approach to be easier to implement than the maximum-likelihood method, we encountered many issues in practice when fitting an ellipse model to the simulated data using Bayesian EIV regression. For instance, the credible interval for the angle of rotation, α, does not cover the true value.

The two methods were further tested when we fit an ellipse to the optical fiber data. The maximum-likelihood method was able to produce parameter estimates, but the Bayesian EIV regression approach could not due to lack of convergence of the chains sampled using HMC. However, parameter estimates were obtained using Bayesian EIV regression when the angle of rotation was fixed at -0.55 rad. The failure of the Bayesian method stems from an inability to estimate α given the current model framework. The angle of rotation and the angles θ_i that correspond to each data point are likely not separable in this model formulation, especially when the data are very circular as is the case for the optical fiber data. Thus, a different model formulation is likely required. The maximum-likelihood method proceeds in a different fashion, minimizing the objective function subject to the constraint of the circle model. Imposing a constraint in an intuitive way, while maintaining the error-in-variables structure of the problem, seems much more difficult in a Bayesian framework. More work is needed to develop new models for Bayesian EIV regression.

Another assumption that should be examined in future Bayesian models is the independence of the errors in the x and y values. For most Bayesian EIV regression problems, this assumption makes sense, but for optical fiber data these errors are likely correlated. It is plausible that a large error in the x coordinate suggests a deformation in the fiber at that point, and would likely have a corresponding y value with a large error as well. The model for the data will depend on how the data are acquired. Future work will include formulating a more accurate model of the error structure for our data. Additionally, in the future we will try using more prior distributions that do not assume independence between the center, radius, and angles.

A. Code

A.1 Circle

A.1.1 R code

```
library(rstan)

df = read.table("J1-1E.DAT")

niters=1000
n.chains = 3
in.dat = list(
  N=dim(df)[1],
  x=df$V1-mean(df$V1),
  y=df$V2-mean(df$V2)
)

x0init=0
y0init=0
truexinit = (in.dat$x-x0init)
trueyinit = (in.dat$y-y0init)

angle = atan2(trueyinit,truexinit)

inits = list(
  list(
    r = 60,
    x0 = x0init,
    y0 = y0init,
    sdy = .01,
    sdx = .01,
    theta = atan2(trueyinit,truexinit)
  ),
  list(
    r = 60,
    x0 = x0init,
    y0 = y0init,
    sdy = .01,
    sdx = .01,
    theta = atan2(trueyinit,truexinit)
  ),
  list(
    r = 60,
    x0 = x0init,
    y0 = y0init,
    sdy = .01,
    sdx = .01,
    theta = atan2(trueyinit,truexinit)
  )
)


fit = stan(
  file = "circle.stan",
  data = in.dat,
  init = inits,
  iter = niters,
  warmup = floor(niters/2),
  chains = n.chains,
  control = list(adapt_delta = .8)
)
```

A.1.2 Stan code

```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
}

parameters {
  real<lower=0> sdy;
  real<lower=0> sdx;

  real x0;
  real y0;

  real<lower = 0> r;

  vector<lower=-pi(), upper = pi()>[N] theta;
}

transformed parameters {
  vector[N] x_tran;
  vector[N] y_tran;

  for(i in 1:N){
    x_tran[i] = x0 + r * cos(theta[i]);
    y_tran[i] = y0 + r * sin(theta[i]);
  }
}

model {
  r ~ normal(60, 10);

  x0 ~ normal(0,1);
  y0 ~ normal(0,1);
  sdy ~ gamma(2,50);
  sdx ~ gamma(2,50);

  theta ~ von_mises(0,0.1);

  x~normal(x_tran,sdx);
  y~normal(y_tran,sdy);
}
```

A.2 Ellipse

A.2.1 R code

```
library(rstan)

######################## Simulate the data
set.seed(4)

n=1000

x0 = 71
y0 = 74
alphasim = -.55 #couterclockwise rotation
rx= 62
ry = 50

theta = seq(-pi+.00001,pi-.00001,length.out = n)

Xi = x0 + rx*cos(theta)*cos(alphasim) - ry*sin(theta)*sin(alphasim)

Yi = y0 + ry*sin(theta)*cos(alphasim) + rx*cos(theta)*sin(alphasim)

df=data.frame(V1=numeric(n))

sdx_true = .05
sdy_true = .07
df$V1=rnorm(length(Xi),Xi,sdx_true)
df$V2=rnorm(length(Yi),Yi,sdy_true)

######################## Stan settings
niters=600
n.chains = 3
in.dat = list(
  N=dim(df)[1],
  x=df$V1-mean(df$V1),
  y=df$V2-mean(df$V2)
)

# Initial values
r=sqrt(in.dat$x^2+in.dat$y^2)
alphainit=atan2(in.dat$y[which.max(r)],in.dat$x[which.max(r)])
possibleAlphas=c(alphainit,alphainit+pi/2,alphainit+pi,
                 alphainit-pi/2,alphainit-pi)

myAlphaInit=possibleAlphas[possibleAlphas<0 & possibleAlphas>-pi/2 ]
```

```
truexinit = (in.dat$x-0)*cos(-myAlphaInit)-(in.dat$y-0)*sin(-myAlphaInit)
trueyinit = (in.dat$y-0)*cos(-myAlphaInit)+(in.dat$x-0)*sin(-myAlphaInit)

rxinit=sum(abs(range(truexinit)))/2
ryinit=sum(abs(range(trueyinit)))/2

angle=atan2(rxinit*trueyinit,ryinit*truexinit)

inits = list(
  list(
    rx = rxinit,
    ry = ryinit,
    x0 = 0,
    y0 = 0,
    sdy =.01,
    sdx =.01,
    alpha = myAlphaInit,
    theta = angle
  ),
  list(
    rx = rxinit,
    ry = ryinit,
    x0 = 0,
    y0 = 0,
    sdy =.01,
    sdx =.01,
    alpha = myAlphaInit,
    theta = angle
  ),
  list(
    rx = rxinit,
    ry = ryinit,
    x0 = 0,
    y0 = 0,
    sdy =.01,
    sdx =.01,
    alpha = myAlphaInit,
    theta = angle
  )
)

fit = stan(
  file = "ellipse_VM.stan",
  data = in.dat,
  init = inits,
  iter = niters,
  warmup = floor(niters/2),
  chains = n.chains,
  control = list(adapt_delta = .85,max_treedepth=11)
)
```

A.2.2 Stan code

```
data {
  int<lower=0> N;
  vector[N] x;
  vector[N] y;
}

parameters {
  real<lower=-pi()/2, upper = 0> alpha;
  real<lower=0> sdy;
  real<lower=0> sdx;

  real x0;
  real y0;

  real<lower = 0> rx;
  real<lower = 0> ry;

  vector<lower=-pi(), upper = pi()>[N] theta;
}

transformed parameters {
  vector[N] x_temp;
  vector[N] y_temp;

  vector[N] x_tran;
  vector[N] y_tran;

  for(i in 1:N){
    x_temp[i] = rx * cos(theta[i]);
    y_temp[i] = ry * sin(theta[i]);

    x_tran[i] = x0 + x_temp[i] * cos(alpha) - y_temp[i] * sin(alpha);
    y_tran[i] = y0 + y_temp[i] * cos(alpha) + x_temp[i] * sin(alpha);
  }
}

model {
  rx ~ normal(60, 10);
  ry ~ normal(60, 10);

  x0 ~ normal(0,1);
  y0 ~ normal(0,1);

  sdy ~ gamma(2,50);
  sdx ~ gamma(2,50);

  alpha ~ uniform(-pi()/2,0);

  theta ~ von_mises(0,0.1);

  x ~ normal(x_tran,sdx);
  y ~ normal(y_tran,sdy);
}
```

Splett, Jolene D.; Jimenez, Felix; Koepke, Amanda. "Estimating the Parameters of Circles and Ellipses Using Orthogonal Distance Regression and Bayesian Errors-in-Variables Regression." Paper presented at 2019 Joint Statistical Meetings, Denver, CO, US. July 28, 2019 - August 01, 2019.

Acknowledgements

The authors would like to thank their colleagues: Adam Pintar, Blaza Toman, Kevin Coakley, Mike Frey, and Antonio Possolo for their very helpful comments.

REFERENCES

P. T. Boggs, R. H. Byrd, J. E. Rogers, and R. B. Schnabel. *User's Reference Guide for ODRPACK Version 2.01 Software for Weighted Orthogonal Distance Regression*. NIST Internal Report 4834, 1992.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

N. Chernov. *Circular and Linear Regression Fitting Circles and Lines by Least Squares*. CRC Press, Boca Raton, FL, 2011.

W. A. Fuller. *Measurement Error Models*. John Wiley and Sons, 1987.

A. Keksel, F. Ströer, and J. Seewig. Bayesian approach for circle fitting including prior knowledge. *Surface Topography: Metrology and Properties*, 6(3):035002, 2018.

T. Lafarge and A. Possolo. The NIST Uncertainty Machine. *NCLSI: Measure The Journal of Measurement Science*, 10(3):20–27, 2015.

L. Mamileti, C.-M. Wang, M. Young, and D. F. Vecchia. Optical fiber geometry by gray-scale analysis with robust regression. *Applied Optics*, 31(21):4182–4185, 1992.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley and Sons, Ltd., New York, NY, 2000.

Stan Development Team. RStan: the R interface to Stan, 2018. URL http://mc-stan.org/. R package version 2.18.2.

C.-M. Wang, D. F. Vecchia, M. Young, and N. A. Brilliant. Robust regression applied to optical-fiber dimensional quality control. *Technometrics*, 39(1):25–33, 1997.

M. Werman and D. Keren. A Bayesian method for fitting parametric and nonparametric models to noisy data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(5):528–534, 2001.

HOW TO USE AND HOW NOT TO USE CERTIFIED REFERENCE MATERIALS IN INDUSTRIAL CHEMICAL METROLOGY LABORATORIES

John R. Sieber
Chemical Sciences Division, National Institute of Standards and Technology
Gaithersburg, Maryland

ABSTRACT

As a producer of certified reference materials (CRMs), NIST faces high demand for Standard Reference Materials (SRMs). The demand is exacerbated by wide spread misuse of CRMs. When should one use CRMs? When should one not use CRMs? Must labs always use NIST SRMs? How can labs demonstrate analytical capabilities for their accreditation scopes? Why so many questions? Standards developers, laboratory accreditors, and laboratory staff must be able to understand these topics with respect to quality systems in compliance with ISO/IEC 17025. They must calibrate and validate test methods and document traceability to the International System of Units (SI). Many people working in laboratory accreditation and under the umbrella of a quality system don't fully understand what these things are, let alone the language of chemical metrology. On average, they have little training in analytical chemistry, elemental analysis, and reference material development. It is hoped this paper will impress upon the reader the need for understanding of how CRMs can be best used in the laboratory. This paper provides a brief background on the above problems and then looks at some of the support and reference information provided by NIST to metals and mining industries labs, commercial CRM producers, and accrediting bodies. The concepts and guidance apply broadly to chemical metrology and fundamental analytical chemistry. The paper includes examples (some from XRF) to illustrate concepts.

STATE OF AFFAIRS IN APPLIED CHEMICAL METROLOGY

To put the uses of certified reference materials into a modern perspective, consider that accreditation is now a *de facto* requirement to do business. For business purposes, it began with ISO 9001 more than 30 years ago. ISO is the International Organization for Standardization, which is one of the leading organizations for development of voluntary consensus standards for a wide range of topics, including laboratory organization practices and standard test methods. Laboratories find compelling business reasons to adopt an ISO/IEC 17025 (2017) quality system. ISO/IEC 17034 (2016) and the ISO Guide 30 series are required for organizations that produce certified reference materials, who may also need to comply with ISO/IEC 17025. However, these

Sieber, John. "How to Use and How Not to Use Certified Reference Materials in Industrial Chemical Metrology Laboratories." Paper presented at Denver X-Ray Conference, Lombard, IL, US. August 05, 2019 - August 09, 2019.

documents are difficult to digest, especially when there is little experience within the analytical community. The difficulty is compounded when there is little or no experience or guidance in how and when to best use certified reference materials. During two ASTM International/National Institute of Standards and Technology workshops in 2004 and 2014, it was shown that roughly four of five auditors from accrediting bodies came from the world of physical metrology, not chemical metrology. Unfortunately, these quality system auditors were not equipped to work with laboratories doing analytical chemistry.

With the combination of a low level of analytical chemistry expertise by auditors and management of accrediting bodies and by quality assurance officers at industrial companies, some audit requirements have been implemented that are incorrect for chemistry labs. The staff of labs undergoing audits typically do not know how and when to challenge requirements. They may also be dissuaded from doing so by their quality officer, because there is a desire to avoid jeopardizing the accreditation.

In seeking guidance, lab staff may look to the organizations from which they obtain certified reference materials (CRMs). After all, CRMs are an important and visible tool in the maintenance of quality performance of test methods. Commercial CRM producers are not necessarily equipped to help customers, because they may have a relatively low level of knowledge of the aspects of chemical metrology that apply to laboratory quality systems and uses of CRMs. A recent, informal survey of certificates of analysis issued by commercial reference materials producers uncovered shortcomings including incorrect treatments of uncertainty, mixing of certified and non-certified values in a single reporting table, lack of rigorous statistical approaches, and lack of indications of overall understanding of the fundamental concepts of CRM development: measurand, traceability, uncertainty, and validation.

Empirical observations appear to show that organizations involved in industrial chemical metrology and the support of industry labs are weak in the expertise needed to perform method validation and to utilize the statistical tools that support demonstration of the quality of test results. Consequently, unnecessary and expensive work is done in the name of quality. As a direct result, NIST SRMs have come to be seen as a panacea to satisfy accreditors and clients. If a lab uses NIST SRMs in calibrations, they are likely to get fewer questions and challenges from auditors, because auditors and the manufacturers' clients believe that is the way to get a quality calibration and to ensure the entire lab operation obtains quality results. This is simply not true, nor is it possible in all cases, because NIST cannot provide SRMs for all analytical needs.

PURPOSES FOR CERTIFIED REFERENCE MATERIALS

At NIST, CRMs are seen as providing higher order references with values for the true amounts of constituents and properties. Certified values are estimates of the true values, because the values are determined based on testing sufficient to elucidate biases in individual test methods, or at least among the methods (May, et al., 2000). Typically, that involves using multiple, independent test methods and examining how well the results from those methods agree. Independent test methods do not share the measurement method, and preferably, they do not share methods of sample preparation. In addition, quality assurance materials are analyzed by each method to critically evaluate the individual method.

In the ISO/IEC 17034 system, it is allowed to certify values for method-dependent definitions of a measurand. The measurand is simply what has been measured. For example, an analytical instrument may measure X-ray photons, for the example of X-ray fluorescence spectrometry (XRF), but the test method determines the mass fraction of an element, which is the measurand. A method-dependent value for a measurand would be obtained if only one test method is used for the assignment of the certified value for the reference material. Using a single test method, it is not possible to test for bias in a way that ensures the test result is an estimate of the true value.

Certified reference materials are primarily tools for validation of test methods, the idea is to use a CRM as if it is a typical sample. See the right side of Figure 1, which maps out how a laboratory result can be traced back to the SI. The test results are evaluated to see if they agree with the certified value. It is recognized that CRMs are also frequently used as calibration standards. See the left side of Figure 1. CRM use is unavoidable, and even necessary when using test methods such as XRF and spark optical emission spectrometry (Spark-OES). These methods are frequently used with solid form materials, e.g. alloys. Both techniques are typically used to determine 10 elements or more in alloys, and both typically require tens of calibration standards to fully calibrate such a method.

CRMs provide values with traceability links to units of the International System (SI), including the kilogram, the mole, and other common units. The SI is not the only standards system that may be used, but it is the world standard system for chemical metrology.

CRMs are often artifacts, that is they are exemplars of real materials with extraordinary homogeneity, often much better than is typically produced for the normal, industrial use of a material. However, many materials do have as-manufactured homogeneity that is fit for the purpose of certification.

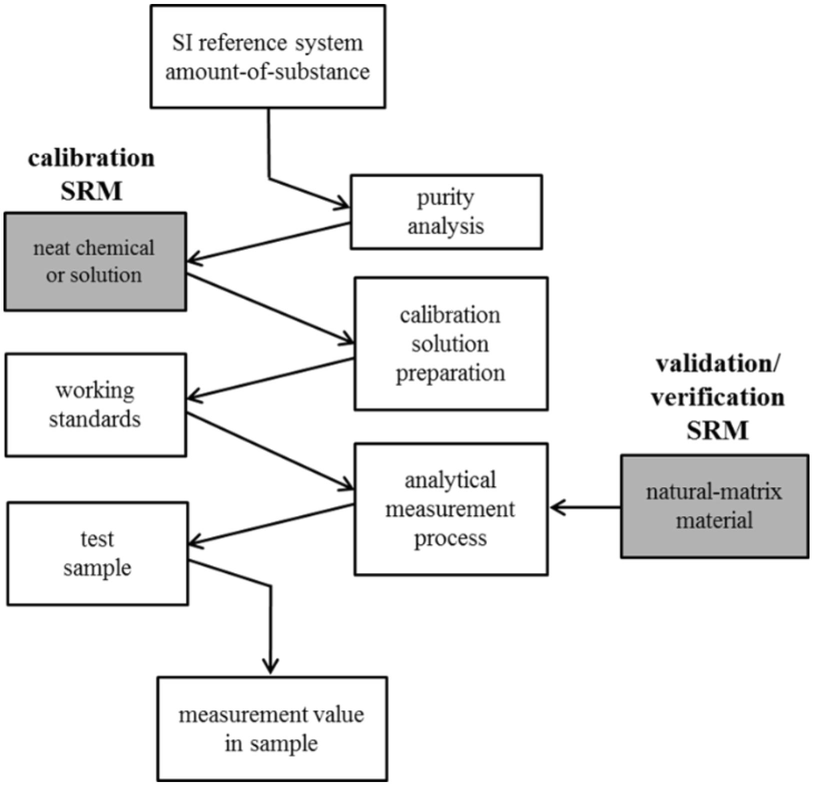


Figure 1. Traceability chain in chemical metrology showing points at which CRMs are used for calibration or validation. The SI is the International System of Units for metrology.

For reasons that will be explained below, one should know that CRMs are rare, expensive, and require long development times. NIST cannot guarantee the continued or continuous supply of any SRM. It may go out of stock with no notice to customers, and it may not be renewed. It may be put on sales restriction due to technical issues, e.g. stability testing, which may result in withdrawal of the material. NIST has a complex system for evaluation of the fitness of materials for their purpose and for the needs and potential impacts for SRMs. SRM users are a critical source of feedback to the NIST Quality System for Measurement Services. Their feedback helps define the needs for reference materials and reference methods in areas of both existing and emerging technologies.

PURPOSES FOR REFERENCE MATERIALS OF OTHER TYPES

The term reference material, in the global lexicon, covers all types of reference materials with CRMs being a subset. At NIST, the two categories are separate and are called Standard Reference Materials (SRMs) and reference materials (RMs). SRMs have at least one certified value based on the NIST criteria for certification and RMs are materials that have no certified values. When there are certified and non-certified values in the same COA, it is required by ISO Guide 31 (2015) to keep them separate. The goal is to prevent confusion by users of the CRM.

Note that in-house RMs can be developed in a single lab or developed by a third party under contract to the lab needing them. There are a number of reasons to use a RM instead of a CRM. Most reasons are based on the amount of information necessary for the intended use and the lower investment needed to develop an RM compared to a CRM.

The first use for an RM may be in process control, that is for use of control charts or control limits to maintain statistical control of a measurement process, i.e. a test method. For this purpose, the RM need not have a certified value. One can simply obtain a value for the measurand from a freshly implemented and validated calibration of the process to be controlled. Multiple measurements provide a mean value as the nominal target value, and the repeatability standard deviation is used to calculate warning and control limits. The primary need is to have a material in good supply and with a high level of homogeneity fit for the purpose, meaning the standard deviation from measurement to measurement, day to day and piece to piece is low enough to provide the required statistical control and not inflate the contribution of process control to the overall uncertainty of test results.

The same requirements apply to using a RM for drift correction. The need is for a homogeneous material that provides a stable signal measurable with high precision. Drift correction options may use more than one material to control drift of sensitivity and baseline. There is no need to certify the material. Simply obtain a value for the signal at time zero and compare that to future measurements to perform drift correction, if it is necessary.

A third use for an RM is instrument conditioning, that is measurements of a material to warm up and to stabilize the instrument response for a particular matrix. It is surprising how often labs will use expensive CRMs for this purpose. An excellent example comes from inert gas fusion and combustion test methods. People have been known to take a $1,000 bottle of an SRM and run six, eight, or even 10 samples just for this purpose. That is a great waste, because they don't need a certified material. They only need something with sufficient homogeneity to demonstrate the instrument is working correctly with the repeatability precision necessary for quantitative determinations of the species being measured. In the metals industry, ASTM International Committee E01 on Analytical Chemistry of Metals, Ores and Related Materials has made progress toward ending this behavior by including statements in the standard test methods that this should not be done.

Calibration is another common use for RMs. There is some debate about that, especially among quality system accreditors and advisory panels, who subscribe to the idea that the only way to get accurate calibrations is to use CRMs as the calibration standards. Many of them also believe the valid range of a calibration extends just from the lowest value from a CRM used as a calibrant to the highest value from a CRM used as a calibrant, and no farther in either direction. Of course, these beliefs ignore basic analytical chemistry and the concepts of limit of quantification at the low end and acceptable levels of uncertainty at the high end of a calibration curve.

One important part of calibration is evaluation of matrix influence and spectral interferences. Reference materials developed for this purpose may be the only viable tools for evaluating the magnitudes of such effects and calculating correction factors, in the absence of fundamental parameters modeling, such as with XRF methods. It may be difficult, if not impossible, to find CRMs that have the levels and variety of compositions necessary for empirical estimation of corrections for any given material matrix. Perhaps the most expedient way to obtain such materials is development of in-house reference materials. It may also be possible to commission a private sector CRM producer to develop a set of RMs.

ASTM Committee E01 published ASTM E2972 Standard Guide for Production, Testing, and Value Assignment of In-House Reference Materials for Metals, Ores, and Other Related Materials (2015). This standard guide was developed to provide explanations of the types and amounts of information necessary to create an in-house RM.

WHAT TO LOOK FOR IN A CERTIFICATE OF ANALYSIS

This discussion below covers just five categories of information required in a certificate of analysis. A complete list is found in ISO Guide 31. The first category is a description of the material. It is up to the CRM user to decide which materials are the best for their intended purpose. The description explains the composition of the material and gives the unit form and size. The form may be a bottle of chips of metal or powder, or it may be a disk of metal or glass. For bottled material, the mass of the bottle contents is given. For solid forms, the shape and dimensions are given.

The next category is a description of the intended uses of the material. For NIST SRMs designed for chemical metrology, the intended uses are summarized by the following sentence. This SRM is intended primarily for evaluation of methods of analysis for similar materials and for validation of value assignment of in-house RMs. There may be specialized uses, possibly related to specific test methods, validation of results for regulatory purposes, or calibration of special equipment.

Next, and perhaps most important is an explanation of the certified values. There are three topics which must be addressed: (1) definition of measurands, (2) explanation of metrological traceability, and (3) definitions of uncertainty estimates. For the definition of measurands, a typical statement is as follows. The measurands are the mass fractions of the total amounts of the elements in a steel matrix. Other possibilities include measurands that are chemical compounds or chemical states of elements. The matrix may be any type of material from steel and other alloys to plastics, foodstuffs, water, fuel oils, and many more. Metrological traceability states the units system for the certified values, which is discussed further below. Uncertainty estimates are typically defined as expanded uncertainty estimates expressed at a coverage level of approximately 95 % and a more detailed discussion of uncertainty follows the details on traceability.

The period of validity of certification is relatively simple in that it can be an expiration date or indefinitely, which means the assigned values and uncertainty estimates for the material are expected to remain valid for a very long period of time. This option is used when a material is known to be highly stable, for example most metals and alloys, when stored correctly. Because CRM producers must perform stability testing and respond to customer inquiries, all CRMs are monitored for their stability.

The last category of information on this top five list is instructions for storage, handling and use. Users need to know how much of a unit is certified. Perhaps the material is a disk of chill-cast metal, and it may be certified for only the first 10 mm deep from the original test surface. The user will also need to know any instructions for sample preparation. If the assigned values are given in a dry basis, instructions for drying are given. Another example is a warning about potential contamination of a disk of metal when a fresh surface is prepared by grinding. Bottles of powder and liquids must be carefully mixed prior to sampling. Every certificate of analysis should provide a minimum recommended sample quantity for an analysis. This will be explained later under the heading of heterogeneity. Finally, the instructions may explain how to use the uncertainty estimates for comparisons of values and for propagation of uncertainty, for which there is more information later in this paper.

METROLOGICAL TRACEABILITY

It is the responsibility of the CRM issuer to establish metrological traceability (traceability for short) of assigned values to a higher-order reference system such as the SI. Traceability is the association of units with an assigned value. In Figure 1, the measured value at the bottom has units from the SI at the top, when the two are connected by calibrations with uncertainty estimates. In the figure, the traceability chain is the assemblage of actions in the boxes and the arrows connecting those boxes. Here is an example SRM statement: “The certified values are metrologically traceable to the SI derived unit of mass fraction expressed as percent”. For NIST SRMs, this is usually sufficient wording to ensure SRM customers that NIST has done the work needed to establish traceability. Many CRMs from commercial producers have no statement of this kind. Instead, they list CRMs and chemicals used in the certification process, without an explicit statement that traceability was established to a specific SI unit. It is an indication they need more training in this area.

For a testing lab to discuss traceability, they must have first dealt with uncertainty estimates for their results, and then they may wish to state something like the following: “The result value is traceable to the SI unit as realized by NIST through the value for element Xx in Standard Reference Material YYY”. While there may be value in being able to show a link to NIST, the traceability link can be through values provided by any CRM producer as long as that producer can demonstrate traceability to the SI units system.

It is also possible for a lab to achieve metrological traceability to an SI unit without a Certified Reference Material. The simplest way is to use a high-purity material having an assay with stated uncertainty and a balance calibrated to establish traceability to the kilogram. There are discussions and demonstrations of this concept in the published papers of Gotthard Staats (1988,1990,1993).

Traceability can be established by using a CRM as a quality assurance material. A comparison between measurement results may be viewed as a calibration, if the comparison to a reference material is used to check and, if necessary, correct the quantity value and measurement uncertainty attributed to the measured material (JCGM 200:2012).

HETEROGENEITY

Composition variance is a part of the overall uncertainty of the value of a measurand. Heterogeneity of a material is evaluated based on the purposes for which the material is intended. With that knowledge, the decision can be made as to whether the material is fit for the intended purpose. To test heterogeneity, choose a method or methods based on considerations of minimal sample preparation, small quantity tested, low counting uncertainty, instrument stability, and testing *within-unit* variance versus *among-units* variance. In most cases, this testing must be done after the candidate material has been prepared and packaged as units for sale.

One or more quantitative analysis methods may adequately account for material heterogeneity as an effect contributing to one of its components of variance. The CRM developer should plan quantitative analyses with that in mind and balance it against the amount of work requested of analysts. In the final assessment of overall uncertainty, it may or may not be necessary to have a component of uncertainty explicitly stated to inform a user of the material heterogeneity. The decision must be made for every CRM development project. Some CRM producers believe a standard uncertainty component for heterogeneity must be published for every measurand in every certificate of analysis. This is not true; however, the COA should contain a statement about heterogeneity, especially when sampling is strongly affected. For example, the heterogeneity and structure of a material may be such that certain portions should not be sampled or that for a particular test method multiple portions should be tested, and the results averaged.

On the concept of minimum sample quantity, think of choosing small aliquots like using a more powerful microscope. At some magnification, it becomes apparent the material is very heterogeneous. Users of CRMs need to know how small a sample can be taken that still allows the analyst to expect to obtain a single result with a value and an uncertainty estimate comparable to the certified value and its uncertainty estimate. If smaller specimens than the recommended minimum must be used, the user must take multiple samples and calculate the mean for comparison to the certified value.

When heterogeneity among units of a CRM is large, the difference between the true value for amount of measurand in one unit and the certified value may be large, or the difference between any two units may be large. Comparisons among units and of a unit to the certified value become more difficult. Figure 2 represents this concept, where the top, black curve shows the range of values for individual bottles as the width of the superimposed (blue) rectangle. This is an exaggerated case of high unit-to-unit variance. The central point of the upper curve represents the overall certified value with the width of the curve representing the uncertainty interval estimated for the certified value. The bottom (red) curve represents the composition of a single unit of the material with the central point being the true value of the amount of substance in that one unit. The (green) rectangle superimposed on the lower curve represents composition variance within the single unit as being smaller that the upper (blue) *among-units* variance. Greater heterogeneity among CRM units causes more units to be shifted farther from the overall certified value. Some units may have their (red) curve shifted way to the left, and some may be shifted way to the right. Therefore, comparisons between any two units must allow for the probability of larger differences, which makes the CRM *less* useful for high precision comparisons. The overall composition variance must be minimized with the contribution of *among-units* variability being less significant than *within-unit* variability. That is, the lower (green) rectangle for a single unit should be wider than the upper (blue) rectangle, but not so wide as to cause the problem of excessive heterogeneity within each unit. When heterogeneity is sufficiently low overall, a material is described as having homogeneity fit for purpose.

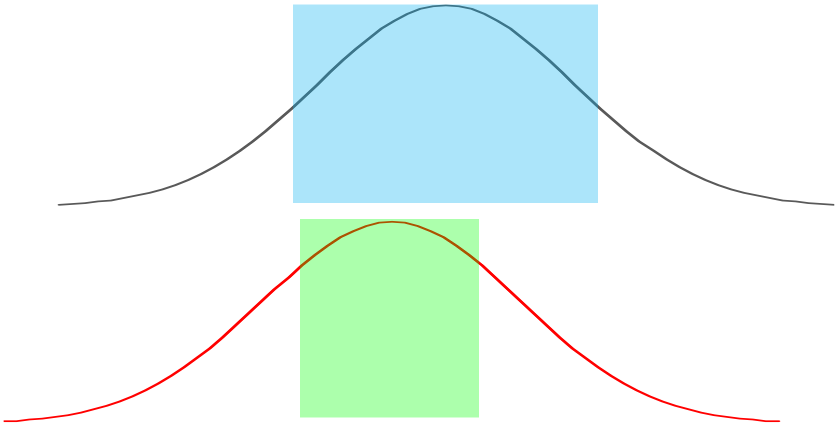


Figure 2. Conceptualization of uncertainty with the top curve representing the certified value for a CRM as the central point of the Gauss curve, the width of the Gauss curve representing 95 % coverage interval, and the top (blue) rectangle representing an exaggerated contribution by overall material heterogeneity. The bottom curve represents the true value for a single CRM unit as the central point of the (red) curve with the curve and (green) rectangle representing the overall and *within-unit* heterogeneity for that unit respectively.

UNCERTAINTY

NIST consensus values and uncertainty estimates for measurands are calculated using statistical methods with many different approaches available, and they can be tailored to each CRM project. Examples of the tools can be seen in the NIST Consensus Builder at https://consensus.nist.gov/, which provides three different approaches for exploration. For those interested in general purpose, statistical tools for evaluation of uncertainty, NIST offers the Uncertainty Machine at https://uncertainty.nist.gov/. Both web tools have online manuals with examples.

In most cases, certified values are accompanied by a symmetric uncertainty interval, expressed as a half-interval, U, with approximately 95 % coverage. That means the full interval, $2U$, is a range within which the true value is expected to be with 95 % confidence. Propagate uncertainty using the combined standard uncertainty, $u_c = U/k$., where k is a coverage factor chosen for the effective degrees of freedom in the evaluation of the certified value. Values for k are given in the certificate of analysis. If for some reason, there is no k value given, the user is advised to set $k = 2$ and to assume the effective degrees of freedom are high enough for that approximation. Older NIST COAs may give just u_c based on variance among collaborator mean results and expert judgement of other components of uncertainty.

In some cases, the uncertainty estimate may be given as an asymmetric interval. Then, the COA will show the certified value accompanied by the range of the interval having approximately 95 % coverage. The decision to provide an asymmetric interval may result from the distributions of values from individual test methods having markedly different widths on either side of the consensus value. In some cases, the consensus value may be so close to either zero or 100 % that the statistics indicate values < 0 or > 100 % are possible. Then, the uncertainty interval will be truncated at zero or 100, making the interval asymmetric. Users can approximate u_c by dividing the range of the coverage interval by four. A conservative alternative is division of the wider side of the range by two. An approach recommended by statisticians is to use a Monte Carlo method for propagation of error, using the actual error distributions representing the uncertainties of the certified values. This method requires fewer assumptions and approximations. To make it possible, the CRM producer must supply files containing the error distributions for the assigned values. NIST SRM 2780a Hard Rock Mine Waste provides that information through a link to the NIST SRM Online Request System:
(https://www-s.nist.gov/srmors/view_datafiles.cfm?srm=2780a).

CRM users must be aware of the great variety of uncertainty definitions to be found in commercial CRM certificates. When in doubt, ask the issuing body to explain their approach and the components of uncertainty included in the estimate. At a bare minimum, the components must include an estimate of repeatability standard deviation from analyses of the material and a standard uncertainty estimate based on the uncertainty of calibration. There may be additional uncertainty components, but without these two things, an uncertainty budget is incomplete.

COMPARING YOUR RESULT TO A CERTIFIED VALUE

The primary purpose of a comparison between a found result and a certified value is to test for bias in the analytical method. ISO Guide 33 (2015) provides an equation for a bias detection limit as shown in Equation 1, where x is the value from each source, u is the combined standard uncertainty for each value, and k is a coverage factor. Most experts agree that $k = 2$ is the most reasonable and convenient choice to approximate 95 % confidence. When the difference between the found result and the certified value is less than this detection limit, there is no evidence for bias in the found result.

$$\left|x_{found} - x_{CRM}\right| \leq k\sqrt{u_{found}^2 + u_{CRM}^2} \qquad (1)$$

The first thing needed for this calculation is an estimate of the uncertainty of the found result. Again, a complete uncertainty budget providing an overall estimate similar in definition to the certificate estimate would be ideal. However, one can begin with repeatability standard deviation and add components for other sources of uncertainty, including calibration, to improve the coverage of the uncertainty of the found result. The more inclusive the uncertainty estimate, the better the information to be gained from the comparison.

The concepts explained above can be illustrated as in Figure 3 in which the four cases are designated: 1) Found result (indicated by X) with no uncertainty either falls outside the certified value uncertainty interval (vertical error bar) (1a), or if the analyst is lucky, it falls inside the interval (1b); 2) Certified value falls within the found value uncertainty interval; 3) Found and Certified intervals overlap and either both values fall inside each other's interval (3a) or both values fall outside each other's interval (3b); and 4) the two intervals do not overlap. For case 1a, there is no evidence the found and certified values agree. For case 4, there is evidence of a bias between the values. For all other cases, there is evidence providing some level of confidence there is agreement. Please note that Eq. 1 may indicate a detected bias in contradiction to Figure 3b, because the standard uncertainty estimates are added in quadrature. It should also be understood that case 1b was referred to as luck, because at NIST the goal is to make the uncertainty estimate as small as possible to show the level of confidence that the certified value is a good estimate of the true value. This goal is in opposition to the goal of including the material heterogeneity variance in the uncertainty estimate, which tends to broaden the coverage interval. In other words, a compromise must be struck to improve confidence in the certified value without hindering the practical utility of the CRM.

EQUIVALENCE OF CRMs AND RMs

The last topic for discussion addresses an example of a faulty accreditation requirement. One well-known requirement in the metals industry is that all CRMs must have been developed under an ISO/IEC 17034 accredited system. While such a situation would be ideal, it causes significant problems in practice. It may be that this requirement resulted from an attempt at universal application of the relatively new ISO/IEC 17034 to all accreditations. Because ISO/IEC 17034 (and its predecessor ISO Guide 34) is a relatively recent phenomenon, there are relatively few CRMs that are compliant out of the hundreds of extant CRMs. The implementation of Guide 34 did not automatically invalidate all existing CRMs at that time. To do that would be to try to argue that none of the CRM producers and national metrology institutes knew how to develop certified reference materials prior to sometime in the 1990s. That simply isn't true. In fact, Guide 34 grew out of efforts by qualified organizations to codify the processes in CRM development for the general benefit of all.

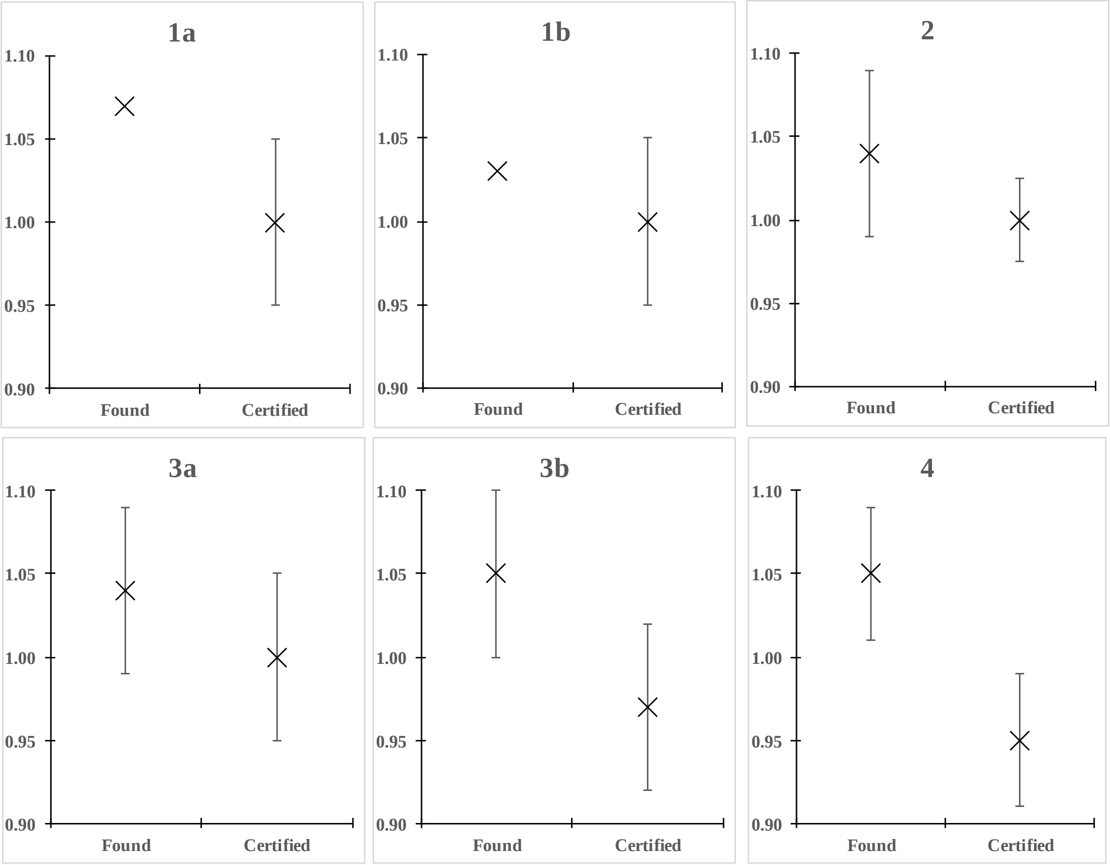


Figure 3. Example cases for comparisons of found results to a certified value and accompanying uncertainty estimate. Case 1a: bias indicated, found value without uncertainty outside certified value coverage interval; Case 1b: no evidence of bias, found value without uncertainty falls inside certified value interval; Case 2: no evidence of bias, uncertainty of found value overlaps certified value; Case 3a: no evidence of bias, found coverage interval overlaps certified value and certified value interval overlaps found value; Case 3b: may or may not be bias, two coverage intervals overlap, but do not cover the values; and Case 4: bias indicated, no overlap of coverage intervals.

A good way to demonstrate that many CRMs in existence prior to the advent of ISO Guide 34 are still good even after the conversion to ISO/IEC 17034 is through the concept of equivalence. Equivalence means the CRMs, test methods or test results from different sources are all fit for the same purpose. The example shown in Figure 4 is calibration of copper in low alloy steel, using CRMs developed from 1965 through 1993. There are 36 CRMs with 33 developed by NIST. The other three were issued by Bureau of Analysed Samples Ltd. (Middlesbrough, England). In the graph, the calibration curve passes through every set of horizontal error bars. That fact and the relatively narrow sizes of the error bars indicate that all 36 CRMs are fit for the purpose of calibration of Cu in steel. In other words, they are all equivalent for that application. Consequently, a subset of the CRMs would be sufficient for the calibration task, and no biases would be expected among results from different subsets of these CRMs.

The concept of demonstration of equivalence has a practical application in testing laboratories. Perhaps a lab has a small collection of CRMs and RMs with two or three each from multiple CRM producers and a number of RMs included. When it can be demonstrated that all of them fit on the same calibration curve within their respective uncertainty estimates, the lab has validation evidence to show to auditors, who may challenge the lab's practices, their choices of CRMs and their inclusion of RMs as calibration standards. Of course, the RMs must have defensible uncertainty estimates that can be used as part of the equivalence comparisons. However, they need not all be certified under an ISO/IEC 17034 compliant quality system.

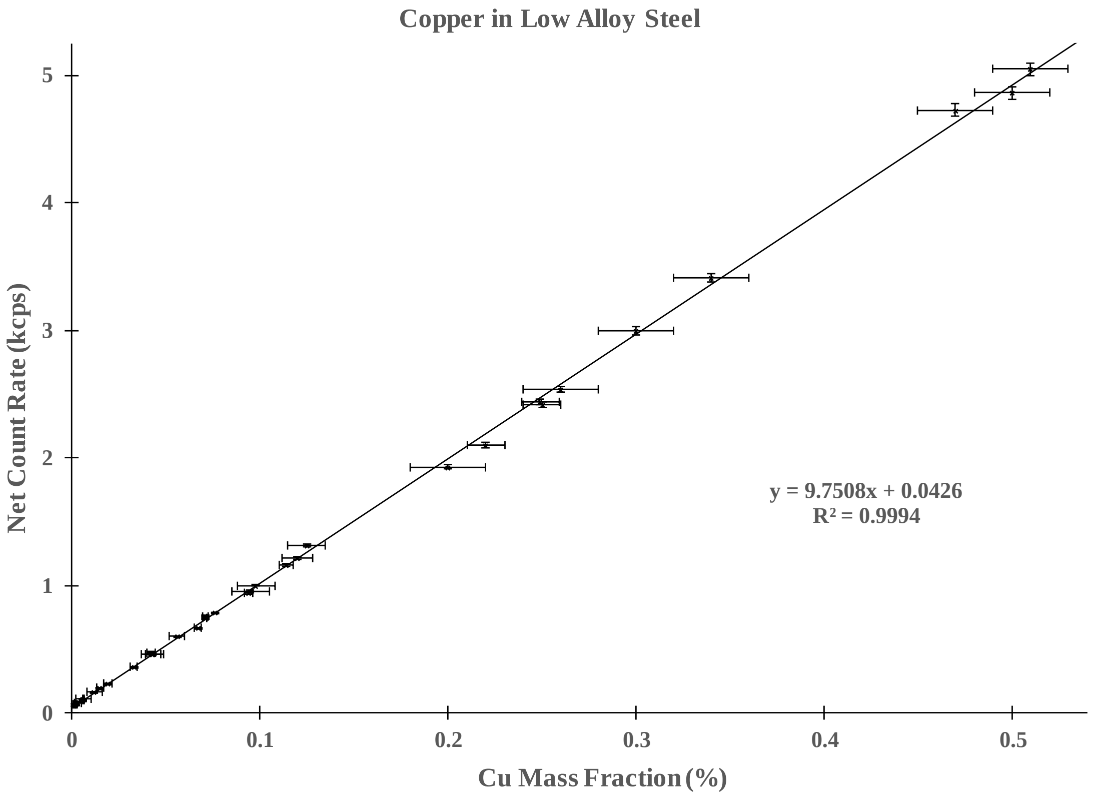


Figure 4. Calibration of a wavelength dispersive XRF spectrometer for copper in low alloy steel, using 33 NIST SRMs and three materials from Bureau of Analysed Samples. Horizontal error bars equal the published uncertainty estimate from each certificate of analysis. Vertical error bars are X-ray counting standard uncertainty.

CONCLUSION

This manuscript is intended to teach a broad audience some fundamentals of certified reference material development, documentation, and use. It began by setting the stage with a description of the state of affairs in industrial chemical metrology in which there is a strong need for education in basic analytical chemistry and in method validation. The discussion covered the primary purposes for CRMs and contrasted them to uses better suited to non-certified RMs. Then, fundamentals of CRM documentation were described to explain what a CRM user needs to ascertain from a certificate of analysis. Of those fundamentals, metrological traceability, heterogeneity and uncertainty were explained in greater detail from the NIST perspective. The paper finished with two examples of how to interpret measured results for CRMs, including testing for bias between found results and certified values and demonstration of equivalence of reference materials. Following the references list below, the reader will find a list of recommended reading on some of the topics discussed herein.

SRM development and other support provided by NIST is done with user needs in mind. NIST welcomes feedback and requests for support. The most direct way to provide feedback is by contacting the NIST Technical Contact for the individual SRM of interest, or one that is similar to the composition needed. For example, see SRM 8k (NIST, 2017) for which this author is Technical Contact. Questions of a more general nature can be sent to srms@nist.gov, or one may visit the Standard Reference Materials website at www.nist.gov/srm/.

DISCLAIMER

Certain commercial equipment, instrumentation, or materials are identified in this document to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

ASTM E2972 − 15 "Standard Guide for Production, Testing, and Value Assignment of In-House Reference Materials for Metals, Ores, and Other Related Materials"; Annu. Book of ASTM Stand.; vol. 03.05; ASTM International, West Conshohocken, PA, 2015, DOI: 10.1520/E2972-15.

ASTM Workshop on the Impact of Accreditation on Reference Materials and Method Development in Analytical Laboratories Supporting the Metals and Ores Industries; Gaithersburg, MD, USA; 19 May 2014; https://www.nist.gov/news-events/events/2014/05/astm-workshop-impact-accreditation-reference-materials-and-method (accessed Sept. 2019)

ISO/IEC 17025:2017 "General Requirements for the Competence of Calibration and Testing Laboratories"; International Organization for Standardization, Geneva.

ISO/IEC 17034:2016 "General Requirements for the Competence of Reference Material Producers"; International Organization for Standardization, Geneva.

ISO Guide 31:2015 "Reference Materials – Contents of Certificates and Labels"; International Organization for Standardization, Geneva.

ISO Guide 33:2015 "Uses of Reference Materials – Good Practice in Using Reference Materials"; International Organization for Standardization, Geneva.

JCGM 200:2012 "International vocabulary of metrology – Basic and general concepts and associated terms"; (VIM), 3rd edition, 2008 version with minor corrections; International Bureau of Weights and Measures, Paris; https://www.bipm.org/utils/common/documents/jcgm/JCGM_200_2012.pdf (accessed Sep. 2019).

May, W., Parris, R., Beck, C., Fassett, J., Greenberg, R., Guenther, F., Kramer, G., Wise, S., Gills, T., Colbert, J., Gettings, R., MacDonald, B.,NIST Spec. Pub. 260-136 "Definitions of Terms and Modes Used at NIST for Value-Assignment of Reference Materials for Chemical Measurements" (2000).

Certificate of Analysis "SRM 8k - Bessemer Steel (Simulated) 0.1 % Carbon (chip form)"; 19 July 2017, https://www-s.nist.gov/srmors/view_detail.cfm?srm=8K (accessed Sept. 2019).

Staats, G., "On the role of pure oxide materials for primary calibration and validation in inorganic bulk analysis"; Fresenius J. Anal. Chem. 336, 132 (1990).

Staats, G., "Computer aided conversion of repeatability into trueness as applied to inorganic analytical chemistry"; Fresenius Z. Anal. Chem. 330, 469 (1988).

Staats, G., Strieder, S., "Validation of Sulphide-Bearing Certified Reference Samples by Primary Calibration as Applied to X-Ray Fluorescence"; X-Ray Spectrom. 22(3), 132 (1993).

Recommended reading:

ASTM E2857 – 11 (Reapproved 2016) "Standard Guide for Validating Analytical Methods"; Annu. Book of ASTM Stand.; vol. 03.05; ASTM International, West Conshohocken, PA; DOI: 10.1520/E2857–11R16.

JCGM 100:2008(E) "Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement (GUM 1995 with Minor Corrections)"; International Bureau of Weights and Measures, Paris; https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf (accessed Sep. 2019).

JCGM 101:2008 "Evaluation of Measurement Data – Supplement 1 to the "Guide to the Expression of Uncertainty in Measurement" - Propagation of Distributions Using a Monte Carlo Method"; International Bureau of Weights and Measures, Paris; https://www.bipm.org/utils/common/documents/jcgm/JCGM_101_2008_E.pdf (accessed Sep. 2019).

Duewer, D.L., Lippa, K.A., Rukhin, A.L., Sharpless, K.E.; NIST Spec. Pub. 260-181 "The ABCs of Using Standard Reference Materials in the Analysis of Foods and Dietary Supplements: A Practical Guide"; available at https://www.nist.gov/sites/default/files/documents/srm/SP260-181r1-2.pdf (accessed Aug. 2019).

Taylor, B.N., Kuyatt, C.E.; NIST Technical Note 1297 “Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results”; available at https://physics.nist.gov/cuu/Uncertainty/index.html (accessed Aug. 2019).

40th AIVC - 8th TightVent & 6th venticool Conference, 2019

Techniques to Estimate Commercial Building Infiltration Rates

Andrew Persily*, Lisa Ng, W. Stuart Dols, and Steven Emmerich

National Institute of Standards and Technology
100 Bureau Drive MS8600
Gaithersburg, Maryland USA
andyp@nist.gov

ABSTRACT

The estimation of low-rise, residential building infiltration rates using envelope airtightness values from whole building fan pressurization tests has been the subject of much interest and research for several decades, constituting a major topic of discussion during the early years of the AIVC. A number of empirical and model-based methods were developed, with their predictive accuracy evaluated in field studies around the world. Infiltration estimation methods for residences are now commonly available in guidance documents and employed in whole building energy and indoor air quality modeling. However, the greater complexity of many commercial buildings, including their size, multizone airflow dynamics and the influence of mechanical ventilation systems, makes the estimation of infiltration rates from fan pressurization test results more challenging. As a result, progress on infiltration estimation methods for commercial buildings has been slower than in low-rise residential. This paper reviews methods for estimating commercial building infiltration rates going back to the 1970s. More recent approaches using correlations based on a large number of multizone airflow model simulations are presented as a more feasible approach. Particular attention is given to how energy models have dealt with infiltration estimation in commercial buildings, with a detailed discussion of how the energy analysis program EnergyPlus has considered infiltration. More complex and presumably more accurate methods of accounting for the energy impacts of infiltration in commercial buildings, based on coupled airflow and energy analysis models, are also discussed as they have become more accessible in recent years given the increasing power of personal computing.

KEYWORDS

airflow; commercial buildings; infiltration; modeling

1 INTRODUCTION

Building infiltration rates have important impacts on energy use, indoor air quality (IAQ), and moisture management, and these rates have been studied for more than 80 years (Coleman and Heald, 1940). The importance of infiltration was demonstrated by the creation of the Air Infiltration Centre (AIC), which held its first conference in 1980 focusing on infiltration measurement (Air Infiltration Centre, 1980), and subsequently evolved into the Air Infiltration and Ventilation Centre (AIVC). The term infiltration is used in this paper to describe the airflow into and out of buildings through unintentional leakage in the exterior building envelope due to pressure differences induced by wind, indoor-outdoor temperature differences and the operation of ventilation and other building systems. It is interesting to compare this with the definition promulgated by the AIVC in 1992, i.e., “The uncontrolled inward leakage of outdoor air through cracks, interstices, and other unintentional openings of a building, caused by the pressure effects of the wind and/or the stack effect” (Limb, 1992). This earlier definition mentions only inward flow, although many discussions of infiltration implicitly include outward flow or exfiltration. It doesn’t mention pressure differences induced by building systems, such as exhaust fans, atmospherically-vented combustion appliances and unbalanced ventilation systems. In addition to these pressures, infiltration rates also depend on building envelope airtightness, airtightness of interior partitions and air temperatures within individual

Persily, Andrew K. "40th AIVC - 8th Tight Vent & 6th Venticool Conference, 2019 Techniques to Estimate Commercial Building Infiltration Rates." Paper presented at 40th AIVC - 8th TightVent - 6th Venticool Conference, 2019, Ghent, BE. October 15, 2019 - October 16, 2019.

building zones. The physics of infiltration are well-understood and have been documented in a variety of places, including Dols and Polidoro (2015) and Etheridge and Sandberg (1996).

Much of the early work on infiltration focused on low-rise residential buildings based in part on their relative simplicity compared with commercial and institutional buildings, which are often larger and more complex, and have more zones and more elaborate ventilation systems. For the purposes of this paper, the term commercial building is used to describe commercial and institutional buildings that are used for a variety of purposes (e.g. work, education, healthcare, retail, public assembly) but not as residential living spaces. Multi-family residential buildings are also not included in this discussion. This description of commercial buildings is admittedly quite broad, and it includes buildings of the same size as single-family residences; the key point is that these commercial buildings include many large, complex, and mechanically ventilated buildings. In general, infiltration rates are more challenging to measure in commercial buildings than low-rise residential, again given their multizone layout, larger size and variations in ventilation system layout and operation. These factors often make it difficult to apply single-zone tracer gas measurement techniques as they complicate achieving uniform tracer gas concentration, which is required in using these techniques (ASTM, 2011). Multi-zone tracer methods exist, which can overcome these challenges, but they are much more difficult to apply (Etheridge and Sandberg, 1996) .

Given that infiltration rates in commercial buildings vary with weather and system operation, they need to be measured many times to characterize infiltration in a given building. Such long-term infiltration measurements have only been done in a small number of large commercial buildings (Grot and Persily, 1986). Therefore, in lieu of extensive measurement efforts, methods to estimate commercial building infiltration are needed to support design, energy calculation, IAQ analysis, retrofit planning and other applications. There is a long history of infiltration estimation methods in low-rise residential buildings, going back to the first AIC conference (Kronvall, 1980; Warren and Webb, 1980; Sherman and Grimsrud, 1980). This earlier work focused on methods to estimate infiltration rates from blower door measurements of envelope airtightness and weather conditions. These techniques evolved over time and are widely used and described in practical guidance (ASHRAE, 2017).

In contrast to low-rise residential buildings, the development of infiltration estimation methods for commercial buildings has progressed more slowly. Due to the often greater complexity of commercial buildings, there have been fewer measurements of envelope airtightness and infiltration rates in commercial buildings to support the testing of estimation methods, also contributing to slower advancement. Nevertheless, infiltration estimation methods in commercial buildings have progressed, and this paper outlines their development and summarizes tools currently available to designers, building scientists and other practitioners.

2 EARLY ESTIMATION METHODS

The earliest description of a large building infiltration estimation method known to the present authors was published before the first AIC conference (Shaw and Tamura, 1977). In that work, separate equations developed specifically for tall buildings are presented for infiltration flow rates due to wind and stack effect. The inputs to the wind infiltration component include building height and width, an envelope air leakage coefficient (that could come from a building pressurization test), wind speed and a wind direction adjustment factor. In addition to the leakage coefficient, the stack infiltration equation requires the length of the building perimeter, the indoor and outdoor air temperatures, the height of the neutral pressure level, and a thermal draft coefficient. The authors do not provide guidance on the neutral pressure level, but it can be measured in conjunction with a fan pressurization test. The thermal draft coefficient depends on the airtightness of the exterior walls relative to the interior partitions and captures how the static pressure decreases with height within the building interior, i.e., whether it is linear with height, which would reflect a relatively open interior, or whether there are pressure drops across floors and other interior partitions. The authors note that the value of this coefficient is around 0.8 in office buildings based on the small number they had studied but note the lack of measured values for apartment buildings. The reference also includes an equation to combine the wind and stack infiltration components to estimate the total infiltration rate. In the 42 years since its publishing, Shaw and Tamura (1977) is only cited 34 times in Google Scholar, most of which are publications on infiltration modeling. None of these references appear to have employed the model to predict infiltration rates and compare them with measured values.

The other means of calculating infiltration rates in commercial buildings is multizone airflow modeling, i.e., CONTAM or other software (Dols and Polidoro, 2015; Walton, 1989b; ESRU, 2002; IES, 2019). These models involve a multizone representation of a building and use mass balance analysis to solve for the airflows between all building zones including the outdoors, from which building infiltration rates can be calculated. These models account for all relevant building airflow physics, though their neglect of the conservation of momentum and energy limits their applicability in buildings that are naturally ventilated and with interior air that is not quiescent. Given the need to for many inputs to describe multizone building airflow systems, such models are generally not considered to be readily accessible to practitioners, though they are widely used in the design of smoke control systems (Klote et al., 2012).

3 INFILTRATION ESTIMATION IN ENERGY ANALYSIS

While commercial buildings have always experienced air infiltration through envelope leaks, and early studies showed that the rates were significant (Grot and Persily, 1986), infiltration was often neglected in energy analysis in part due to its perceived complexity. Rather than take this important phenomenon seriously, many simply assumed infiltration was equal to zero. In some cases, at least in the U.S., this assumption was justified by claiming that mechanically ventilated buildings are pressurized, thus eliminating infiltration. This justification was based on the common practice of providing more supply air than return air. In reality, the complexity of multizone building airflow systems results in indoor-outdoor pressure differences being localized phenomena that depend on outdoor wind patterns, building height, and differences between the rate at which ventilation air is delivered to and removed from individual building zones. As a result, infiltration will occur even when more ventilation air is supplied to a building than removed by exhaust unless detailed analyses and control strategies are implemented to control pressure differences across the entire envelope.

Another approach, embodied in the EnergyPlus energy analysis software and other tools, is to use empirical equations for estimating infiltration. These empirical equations were developed from analysis and testing of low-rise residential buildings as noted below, and do not capture the airflow physics of large, multizone or mechanically ventilated buildings. In the case of EnergyPlus, the following equation is available for calculating infiltration rates:

$$\text{Infiltration} = I_{\text{design}} \bullet F_{\text{schedule}} \left[A + B|\Delta T| + C \bullet W_{\text{s}} + D \bullet W_{\text{s}}^2\right] \qquad (1)$$

where I_{design} is defined by EnergyPlus as the "design infiltration rate", which is the airflow through the building envelope under design conditions. Its units are selected by the user and can be h^{-1}, $m^3/s \bullet m^2$ or m^3/s. To apply this infiltration approach in EnergyPlus, a value of I_{design} is assigned to each zone. F_{schedule} is a factor between 0.0 and 1.0 that can be scheduled, typically to account for the impacts of fan operation on infiltration. $|\Delta T|$ is the indoor-outdoor temperature difference in °C, and W_{s} is the wind speed in m/s. A, B, C and D are constants, for which values are suggested in the EnergyPlus Engineering Reference (DOE, 2019). As noted in that reference, this equation is based on measurements in 10 one- and two-story houses, for which 30 infiltration rates were fit to the equation (Coblentz and Achenbach, 1963). The default strategy in EnergyPlus is to assume a constant infiltration rate, i.e., A=1 and B=C=D=0. However, this approach does not reflect known dependencies of infiltration on outdoor weather and HVAC system operation. The EnergyPlus Engineering Reference provides values of A, B, C and D based on two energy analysis programs that preceded EnergyPlus, BLAST and DOE-2. No references are provided for these values, but they are presumably based on studies in low-rise, residential building as there were no studies of infiltration in commercial buildings available when these two predecessor programs were developed. The EnergyPlus Engineering Reference also includes two other empirical infiltration models developed for low-rise residential buildings, i.e., the Sherman-Grimsrud and the AIM-2 models described in Chapter 16 of the ASHRAE Fundamentals Handbook (ASHRAE, 2017). While these approaches account for weather effects, they are also based on low-rise residential buildings and do not account for the airflow physics in larger buildings and in buildings with the more complex mechanical ventilation systems typical of commercial buildings.

Gowri et al. (2009) proposed a method to account for infiltration in commercial buildings that was developed using a square medium-size office building and a building envelope airtightness value, such as can be obtained through pressurization testing. Assuming a constant indoor-outdoor pressure difference of 4 Pa, Gowri calculated an infiltration rate using an approach that accounts for wind but not temperature effects, despite their known importance in taller buildings and colder climates. Gowri recommends that this infiltration rate be multiplied by a wind speed adjustment and by a factor of 0.25 when the HVAC system is on and 1.0 when the system is off. Overall, the method greatly oversimplifies the dependence of infiltration on building envelope airtightness, weather, and HVAC system operation.

EnergyPlus also has the ability to perform multizone airflow analysis, as embodied in the CONTAM model discussed above, using the EnergyPlus Airflow Network model (DOE, 2019). This model is based on a predecessor to CONTAM referred to as AIRNET (Walton, 1989a) and an earlier version of CONTAM (Walton and Dols, 2003). It is worth noting that while CONTAM has evolved considerably in the intervening years, the EnergyPlus Airflow Network model does not incorporate all of those improvements.

Han et al. (2015) compared the use of various infiltration estimation methods in improving the accuracy of building energy simulations. One method used the EnergyPlus Airflow Network model with three infiltration levels (leaky, medium and tight) as defined via DesignBuilder, which is a graphical front-end to EnergyPlus. The other methods utilized various means to establish monthly and annually averaged wind pressures including the use of an AIVC database of wind tunnel measurements and the use of computational fluid dynamics (CFD) to model the building exterior. For this case study, the CFD-based methods resulted in energy predictions that more closely matched utility bills.

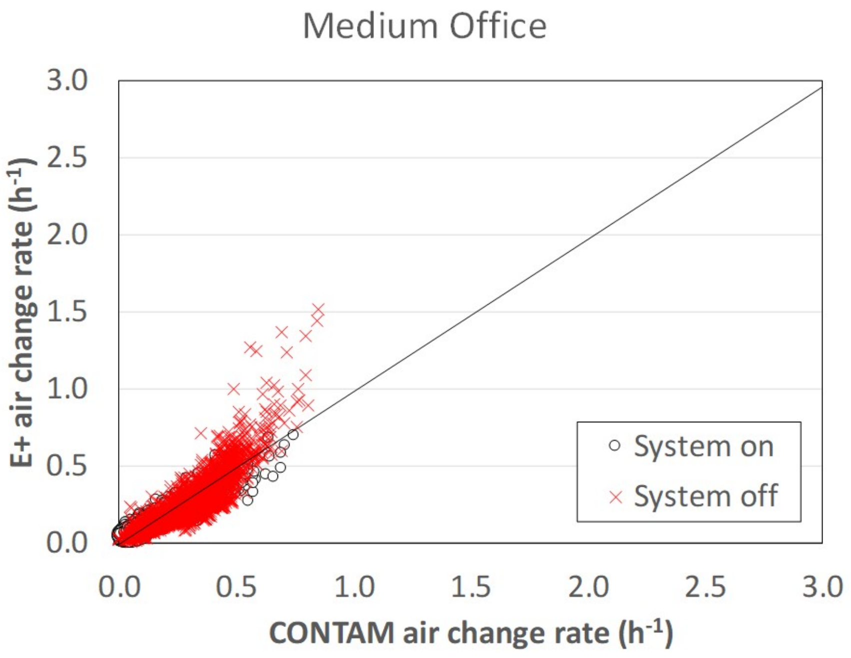
4 MORE RECENT ADVANCES

As building energy use has become of increasing interest over the years and as envelope insulation levels have increased, there has been more recognition of the importance of infiltration, including the need to control air leakage and to reliably estimate infiltration rates. These changes were reflected in the inclusion of requirements for continuous air barriers and envelope airtightness testing in energy efficiency standards such as ANSI/ASHRAE/IES Standard 90.1 (ASHRAE, 2016). As a result, new approaches to estimating infiltration were developed and are becoming more widely applied. This section describes one such approach, which uses values of the coefficients in Equation (1) developed specifically for commercial buildings, as well as a second approach using coupled energy and airflow analysis methods.

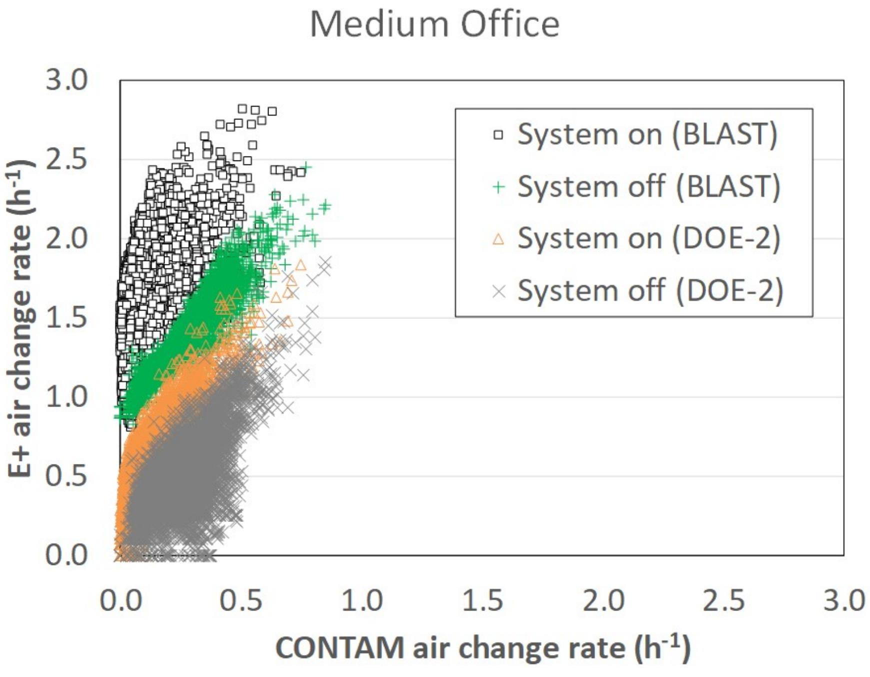
4.1 Large Building Infiltration Correlations

Values of the coefficients in Equation (1) for commercial buildings were recently developed by NIST. These coefficients were identified from investigations of the relationships between infiltration rates calculated using multizone airflow models, weather conditions, and building characteristics, including envelope airtightness and HVAC system operation. These relationships were developed using CONTAM models of the EnergyPlus models of 16 DOE commercial reference buildings (DOE, 2011; Goel et al., 2014). For each of these 16 buildings, different versions of the EnergyPlus models exist that correspond to different versions of ASHRAE Standard 90.1. In the ASHRAE 90.1-2004 versions, infiltration was modeled as 100 % of the design value when the HVAC system was off and 25 % (or 50 %) of that value when the HVAC system was on using the F_{schedule} term in Equation (1). Further, A=1 and B=C=D=0 in Equation (1), which means the weather dependence of infiltration was ignored. In the ASHRAE 90.1-2013 versions of the prototype building models, infiltration was modeled with the same F_{schedule} values but with C=0.224 based on a study by Gowri et al. (2009), which as noted earlier ignores the dependence of infiltration on indoor-outdoor temperature differences. These estimation methods are limited in that they do not fully account for weather effects and greatly oversimplify the impacts of system operation on infiltration.

In NIST's effort to develop coefficients for Equation (1), simulations were performed using NIST-developed CONTAM models of the DOE reference buildings to generate whole building infiltrations rates for a range of weather conditions with the ventilation fans on and off. Correlations were performed to fit these predicted infiltration rates to weather data, i.e., outdoor temperature and wind speed, to estimate the constants in Equation (1). For the ASHRAE 90.1-2004 versions of the building models, correlations were performed for 7 of the 16 reference models using Chicago weather (Ng et al., 2015). Compared with the assumption of constant infiltration in the reference building models, the correlation-based infiltration estimates agreed 60 % better on average with the CONTAM predictions. An example of the agreement between these infiltration correlations and the CONTAM predictions is shown for the Medium Office in Fig. 1a. The values for A, B and D for these 7 buildings were then correlated with building height, exterior surface area to volume ratio, and net system flow (i.e., design supply air minus design return air minus mechanical exhaust air) normalized by exterior surface area in order to predict these coefficients for any given building based on these three parameters (height, surface to volume ratio and net system flow). This generalized method resulted in an average improvement of 50 % when compared to the constant infiltration rates in the reference building models. Figure 1b compares the predictions using the suggested constants from DOE-2 and BLAST in the EnergyPlus Engineering Reference document, which results in much poorer agreement with the CONTAM predictions than seen using the correlation approach.



(a)



(b)

Fig. 1 EnergyPlus infiltration rates vs. CONTAM infiltration rates for Medium Office building using (a) NIST correlations and (b) DOE-2 and BLAST coefficients

In order to make these correlation-based estimation methods more accessible, they were incorporated into an Open Studio Measure called “Adding infiltration correlated for weather and building characteristics” (https://bcl.nrel.gov/node/83101). When the ASHRAE 90.1-2013 versions of the prototype building models were released, NIST updated the corresponding CONTAM models and expanded the infiltration correlations to eight cities (Ng et al., 2018). This work is continuing with the inclusion of more building types.

4.2 Energy-Airflow Model Coupling

Airflow and heat transfer are closely coupled processes in buildings, particularly in multizone buildings and in applications of natural ventilation. Basically, airflow models need zonal air temperatures to calculate airflows, and energy models need these airflows to calculate zonal temperatures. These calculations have historically been handled by separate software tools, with the inputs of each tool entered by the user “manually.” However, increased computer power and more widespread application of building simulation in design is making both building energy analysis and airflow modeling more accessible, as well as enabling more direct coupling between these two types of modeling tools. The CONTAM-based methods employed by Han et al. (2015) utilized loose coupling between the energy and airflow analysis software, i.e., the outputs from one program were used as inputs to another but not during runtime. More recent advancements have enabled tight coupling or co-simulation between CONTAM and energy modeling tools, i.e., EnergyPlus and TRNSYS (Dols et al., 2015; Dols et al., 2016). Co-simulation allows for the runtime exchange of simulation results between energy and airflow programs at each time step. Currently available coupling allows the complex interaction between building-specific thermal, wind and system effects to be addressed simultaneously.

5 SUMMARY AND CONCLUSIONS

The estimation of infiltration rates in commercial and institutional buildings is more complex than in low-rise residential buildings given the larger size and multi-zone configuration of these buildings, the complexity of their HVAC systems and the lack of infiltration and airtightness measurements to validate estimation methods. As a result, there has been slower progress in infiltration estimation compared with low-rise residential buildings. This paper described the history of infiltration estimation in commercial buildings, with a focus on how energy analysis tools (i.e., EnergyPlus) treat infiltration. At this time, there are basically three options for estimating infiltration rates in commercial buildings: multizone modeling, empirical formulas (i.e., Equation 1) in which the coefficients are all non-zero and have a sound technical basis, and coupled energy-airflow modeling (which is essential an application of multizone modeling). While there has been progress in these estimation methods, it is unclear how simple such methods can be given the complexity of airflow and pressure in large, multizone buildings. Nevertheless, infiltration rates are needed for commercial building energy and IAQ analyses or other applications, and if an estimated rate is to be used it must be based on a sound technical approach and data, which need to be reported along with the estimate.

6 ACKNOWLEDGEMENTS

The authors express their appreciation to Sara Persily of BCER Engineering for her helpful review.

7 REFERENCES

Air Infiltration Centre (1980) "1st AIC Conference. Air Infiltration Instrumentation and Masuring Techniques".

ASHRAE. 2016. *Energy Standard for Buildings Except Low-Rise Residental*, Atlanta GA, American Society of Heating, Refrigerating and Air-Conditioning Engineers, (ANSI/ASHRAE/IES Standard 90.1-2013).

ASHRAE. 2017. *Fundamentals Handbook,* Atlanta, GA, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

ASTM. 2011. *Standard Test Method for Determining Air Change in a Single Zone by Means of a Tracer Gas Dilution*, West Conshohocken, PA, American Society for Testing and Materials, (ASTM Standard E741-2011).

Coblentz, C.W. and Achenbach, P.R. 1963. Field Measurements of Air Infiltration in Ten Electrically-Heated Houses. *ASHRAE Transactions*, 69, 358-365.

Coleman, E.F. and Heald, R.H. 1940. *Air Infiltration Through Windows*, National Bureau of Standards, Washington, DC, BMS45.

DOE (2011) Commercial Reference Buildings, U.S. Department of Energy, Available from: https://www.energy.gov/eere/buildings/commercial-reference-buildings.

DOE. 2019. *EnergyPlus Version 9.1.0 Documentation. Engineering Reference*, U.S. Department of Energy.

Dols, W.S., Emmerich, S.J. and Polidoro, B.J. 2016. Coupling the Multizone Airflow and Contaminant Transport Software CONTAM with EnergyPlus Using Co-Simulation. *Building Simulation*, 9, 469-479.

Dols, W.S. and Polidoro, B.J. 2015. *CONTAM User Guide and Program Documentation. Version 3.2*, Gaithersburg, MD, National Institute of Standards and Technology, NIST Technical Note 1887.

Dols, W.S., Wang, L., Emmerich, S.J. and Polidoro, B.J. 2015. Development and application of an updated whole-building coupled thermal, airflow and contaminant transport simulation program (TRNSYS/CONTAM). *Journal of Building Performance Simulation*, 8, 327-337.

ESRU. 2002. *The ESP-r System for Building Energy Simulation. User Guide Version 10 Series*, Glasgow UK, University of Strathclyde.

Etheridge, D.W. and Sandberg, M. 1996. *Building ventilation: theory and measurement,* West Sussex, England, John Wiley & Sons, Ltd.

Goel, S., Athalye, R., Wang, W., Zhang, J., Rosenberg, M., Xie, Y., Hart, R. and Mendon, V. 2014. *Enhancements to ASHRAE Standard 90.1 Prototype Building Models.*, Richland Washington, Pacific Northwest National Laboratory.

Gowri, K., Winiarski, D. and Jarnagin, R. 2009. *Infiltration Modeling Guidelines for Commercial Building Energy Analysis*, Pacific Northwest National Laboratory.

Grot, R.A. and Persily, A.K. (1986) Measured Air Infiltration and Ventilation in Eight Federal Office Buildings, In: Treschel, H. R. and Lagus, P. L. (eds) *Measured Air Leakage of Buildings ASTM STP 904*, Philadelphia, PA, American Society for Testing and Materials, 151-183.

Han, G., Srebric, J. and Enache-Pommer, E. 2015. Different modeling strategies of infiltration rates for an office building to improve accuracy of building energy simulations. *Energy and Buildings*, 86, 288-295.

IES (2019) IES Virtual Environment, Integrated Environmental Solutions Ltd.

Klote, J.H., Milke, J.A., Turnbull, P.G., Kashef, A. and Ferreira, M.J. 2012. *Handobok of Smoke Control Engineering,* Atlanta, GA, American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

Kronvall, J. (1980) "Correlating pressurization and infiltration rate data - tests of an heuristic model". In: Proceedings of Air Infiltration Instrumentation and Measuring Techniques, 1st Air Infiltration Centre Conference, Vol. 1, pp. 225-244.

Limb, M.J. 1992. *Air Infiltration and Ventilation Glossary*, Air Infiltration and Ventilation Centre, Coventry, Great Britain., AIVC Technical Note 36.

Ng, L.C., Ojeda Quiles, N., Dols, W.S. and Emmerich, S.J. 2018. Weather correlations to calculate infiltration rates for U. S. commercial building energy models. *Building and Environment*, 127, 47-57.

Ng, L.C., Persily, A.K. and Emmerich, S.J. 2015. Improving infiltration modeling in commercial building energy models. *Energy and Buildings*, 88, 316-323.

Shaw, C.Y. and Tamura, G.T. 1977. The Calculation of Air Infiltration Rates Caused by Wind and Stack Action for Tall Buildings. *ASHRAE Transactions*, 83 (II), 145-157.

Sherman, M.H. and Grimsrud, D.T. (1980) "Measurement of infiltration using fan pressurization and weather data". In: Proceedings of Air Infiltration Instrumentation and Measuring Techniques, 1st Air Infiltration Centre Conference, Vol. 1, pp. 277-322.

Walton, G.N. 1989a. Airflow Network Models for Element-Based Building Airflow Modeling. *ASHRAE Transactions*, 95 (2), 611-620.

Walton, G.N. 1989b. *AIRNET - A Computer Program for Building Airflow Network Modeling*, National Institute of Standards and Technology, Gaithersburg, MD, NISTIR 89-4072.

Walton, G.N. and Dols, W.S. 2003. *CONTAMW 2.1 Supplemental User Guide and Program Documentation*, Gaithersburg, MD, National Institute of Standards and Technology, NISTIR 7049.

Warren, P.R. and Webb, B.C. (1980) "The relationship between tracer gas and pressurization techniques in dwellings". In: Proceedings of Air Infiltration Instrumentation and Measuring Techniques, 1st Air Infiltration Centre Conference, Vol. 1, pp. 245-276.

LIVE QUANTIFICATION OF CELL VIABILITY VIA NEUTRAL RED UPTAKE USING LENS-FREE IMAGING

Brian J. Nablo[‡], Jung-Joon Ahn[‡], Kiran Bhadriraju[†], Jong Muk Lee[‡], and Darwin R. Reyes[†*]

[†] *National Institute of Standards and Technology (NIST), USA*

[‡]*SOL Inc., Korea*

ABSTRACT

We present the quantification of cell viability during neutral red (NR) uptake with a compact lens-free system utilizing two light sources. Conventionally, the NR uptake assay determines cell viability based on the accumulation of NR within lysosomes and is quantified spectrophotometrically after a destructive extraction process. Our NR uptake live imaging system offers *in situ* monitoring of cell viability while accentuating 3D morphologies, thus diversifying the assay's functionality. Combining the low-cost NR uptake assay with the compact architecture of lens-free imaging provides high-throughput, point-of-use, live monitoring of cell viability that eliminates extraction steps and preserves the culture.

KEYWORDS: Lens-free imaging, Cell Viability, 3D Culture, Hepatic Cancer Cells

INTRODUCTION

Determining cell viability for *in vitro* tissue culture is crucial for evaluating cell health, toxicity of compounds, and efficacy of therapeutics. NR is a common and inexpensive supravital stain for assessing cytotoxicity. Healthy cells accumulate NR as the nonionic dye diffuses into an acidified lysosome and becomes protonated. In the standard assay, cell viability is quantified spectrophotometrically as an ensemble average of the culture after destructive extraction of the accumulated dye.[1] Lens-free imaging provides real-time culture monitoring at a lower cost, wider field of view, and smaller footprint for high-throughput, point-of-use, quantitative live-cell imaging. In lens-free imaging, a light source directly illuminates the sample that is within immediate proximity of the CMOS sensor array. The result is a shadow image with the size and resolution of the array. We present the use of lens-free imaging to monitor viability of human liver cancer cells during NR uptake, rather than after extraction.

EXPERIMENTAL

Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the NIST, nor does it imply that the materials or equipment is necessarily the best available for the purpose.

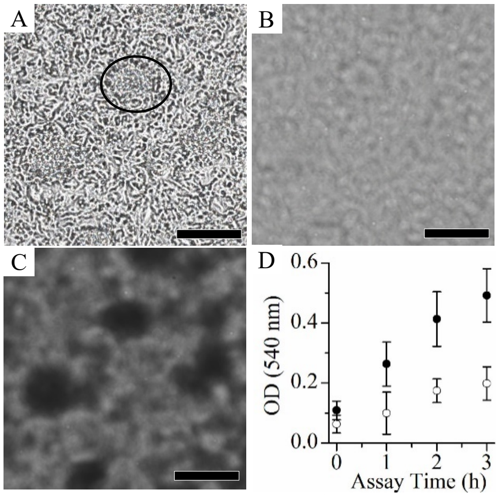


Figure 1. Lens-free analysis of NR uptake by a confluent culture. (A) Pinhole light at 0 h. Collimated light at (B) 0 h and (C) 3 h. (D) NR uptake by clusters (solid) and monolayers (open). Scale bar = 200 µm.

Culture vessels were created by plasma bonding polydimethylsiloxane (PDMS) wells or micropatterns to #1 glass coverslips, then sterilized with 70% ethanol. Flow in microchannels (400 µm wide, 70 µm high, 30 mm long) was controlled with a gentle gravity flow ($\Delta h_{max} = 0.9$ cm). A hybrid cell adhesion matrix (HCAM) was prepared on the surfaces with 50 mg/L fibronectin in Dulbecco's phosphate buffered saline followed by 1 g/L poly(allylamine) in deionized water.

Human hepatoma cells (HepG2) were cultured in Dulbecco's modified eagle medium supplemented with 10% fetal bovine serum (DMEM) at 37 °C and 5% CO_2, harvested at 80% confluency, seeded in culture vessels at a 1:1 vessel:flask surface density and stained within (5 to 6) d. The lens-free system (SOL Inc., KR) has an 8 megapixel CMOS sensor (field-of-view = 10.14 mm^2, spatial resolution = 1.10 µm, pixel size = 1.12 µm). Light sources were filtered at 540±35 nm. The NR uptake assay was performed similar to the standard protocol.[1] After pre-assay imaging, dye-free DMEM was replaced with 40 mg/L NR in DMEM and every hour thereafter for 3 h. Final images were collected with dye-free DMEM. Images were analyzed with

Nablo, Brian; Ahn, Jungjoon; Bhadriraju, Kiran; Lee, Jong; Reyes-Hernandez, Darwin. "Live quantification of cell viability via neutral red uptake using lens-free imaging." Paper presented at The 23rd International Conference on Miniaturized Systems for Chemistry and Life Sciences (µTAS 2019), Basel, CH. October 27, 2019 - October 31, 2019.

Image J[2]. Briefly, pinhole-illuminated images were used to define regions for analysis by subcategorizing cell into “Dense clusters” (circular bodies not in excess of 75 pixels [91 μm^2] or “Spread monolayers” (oblong bodies up to 500 pixels [665 μm^2]). For confluent cultures, areas of consistent cell morphology were marked with circular regions of 1976 pixels (2390 μm^2). Marked regions were transferred to respective collimated-light images to determine the transmittance and calculate absorbance.

RESULTS AND DISCUSSION

The lens-free imaging resolves cells when using a pinhole light source due to cell perimeter scattering light that appear darker and the cell bodies focusing light, frequently saturating the pixel below. This artifact is rendered negligible with a collimated light source (Fig. 1A and B). Thus, imaging occurs in tandem with 1) the pinhole source locating cells and, 2) the collimated source quantifying optical density. Two distinct morphologies are observed herein: 1) spread monolayers, or 2) dense clusters. Under pinhole illumination, dense clusters (Fig. 1A oval) are darker and appear to organize into 3D morphologies similar to multicellular spheroids previously observed on HCAM[3].

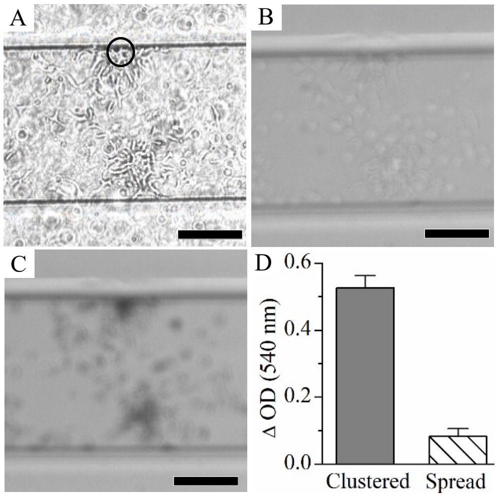


Figure 2. Lens-free analysis of NR uptake by HepG2 cells in a microfluidic channel. (A) Pinhole light at 0 h. Collimated light at (B) 0 h and (C) 3 h. (D) Absorbance of NR in cells. Scale bar 200 μm.

HepG2 cells steadily accumulate NR over the 3 h assay (Fig. 1C and 1D), consistent with the standard NR uptake assay[1]. The dense clusters appear to uptake more NR than spread monolayers. The disparity between morphologies may be due to a difference in pathlength caused by an increase in cell height. Previous work demonstrates that HCAM stimulates HepG2 cells to form 3D cultures[3]. Confocal images of fixed cultures confirm the heights of dense clusters vary from (15 to 35) μm, whereas spread monolayers vary from (4 to 6) μm.

The elimination of the extraction process enables the NR uptake assay to quantify cell viability in microfluidics. Despite the presence of debris within the PDMS, cell perimeters are still identifiable with the pinhole source (Fig. 2A) and NR accumulations is unobstructed with collimated light (Fig. 2B and 2C). Spread cells exhibit a greater cell area while clustering cells (Fig. 2A circle) are more confined and spherical. Clustered cells absorb more light than the spread cells (Fig. 2D), likely due to concentration in lysosomes caused by a reduced cell area.

CONCLUSION

Tandem lens-free imaging with pinhole and collimated light sources enables *in situ* quantification of cell viability and reduces the time to verification for the NR uptake assay. Additionally, NR uptake visually enhances 3D cell morphologies. The reduced footprint of lens-free imaging combined with low-cost supravital stains like NR enables high-throughput assessment of cell viability on a cell-by-cell basis at the point-of-use (e.g., incubator) where traditional microscopes would be disruptive, inhibitory, or costly.

ACKNOWLEDGEMENTS

This work was supported in part by NIST internal funds, NIST grants No. 70NANB17H246 and 70NANB17H292, the N0002519 of the Korea Institute for Advancement of Technology, and the Ministry of Trade, Industry and Energy of the Republic of Korea. This research was performed in part at the NIST CNST.

REFERENCES

[1] Repetto G.; del Peso A.; Zurita J.L., *Nat. Protoc.*, 3, 1125-1131, 2008.
[2] Schneider, C. A.; Rasband, W. S.; Eliceiri, K. W., *Nat. Meth.*, 9, 671-675, 2012.
[3] Bhadriraju, K.; Hong, J.S.; Lund, S.P.; Reyes, D.R., *ACS Biomater Sci Eng.*, 3, 2559-2569, 2017.

CONTACT

* Darwin Reyes; phone: +1-301-975-5466; darwin.reyes@nist.gov

A Graph Database Approach to Wireless IIoT Workcell Performance Evaluation

Richard Candell*, Mohamed Kashef*, Yongkang Liu*, Karl Montgomery*, Sebti Foufou†

*National Institute of Standards and Technology, Gaithersburg, Maryland, USA
Email: {richard.candell, mohamed.kashef, yongkang.liu, karl.montgomery}@nist.gov
† University of Burgundy, Dijon, France, Email: sfoufou@u-bourgogne.fr

***Abstract*—The workcell is considered a main building block of various industrial settings. Hence, it is examined as a primary testing environment for studying wireless communication techniques in factory automation processes. A new testbed was recently designed and developed to facilitate such studies in workcells by replicating various data flows in an emulated production environment. In this paper, an approach to storing and analyzing network performance data from a manufacturing factory workcell is introduced. A robotic testbed was constructed using two collaborative grade robot arms, machine emulators, and wireless communication devices. A graph database approach was implemented to capture network and operational event data among the components within the testbed. A schema is proposed, developed, and elaborated; a database is then populated with events from the testbed, and the resulting graph is presented. Query commands are then presented as a means to examine and analyze network performance and relationships within the components of the network. Additionally, we demonstrate how to extract correlations between receive signal power and network delay within the testbed using the graph database query language. Finally, using the inherently interconnected nature of the graph database, we discuss applying the graph database approach toward examining more complex relationships between the wireless communications network and the operational system.**

***Index Terms*—industrial wireless, factory automation, testbed, measurement, instrumentation, graph database.**

I. Introduction

Wireless communication is a key enabling technology for the modernization of factory workcells. Modern factory workcells are highly-interconnected networked control systems in which various devices interact and collaborate to accomplish complex and adaptive production orders. The workcell often contains mobile robots that collaborate with other robots or human beings. Requirements of the workcell include the incorporation of mobile collaborative robotics with real-time coordination of motion and tool actuation. As such, the communication of sensor and control information must be ultra-reliable and of low-latency to assure trustworthy operation [1]. Due to an increased demand for ease of installation, reduced costs of deployment and maintenance, and flexibility, wired networks are being gradually replaced with wireless networks. This presents a real challenge for networks and control systems. Compared with wired connections, wireless links have their unique advantages in connecting field sensors and actuators with reduced cabling costs and natural support of mobility [2]; however, most current communications systems lack the latency and reliability supports [3] mandated by factory owners [4], [5]. New wireless protocols are being designed to address reliability and latency concerns of real-time systems such as those in manufacturing automation. These new protocols include advancements, such as the Institute of Electrical and Electronics Engineers (IEEE) 802.11ax standard and 5G cellular evolution as defined by the International Mobile Telecommunications-2020 (IMT-2020) Standard. Both of these two standards employ improved diversity techniques for multiple access of devices as well as 1 ms latency and greatly improved reliability [6].

Evaluation of such systems used in manufacturing environments requires not only rigorously analyzing network performance, but also studying the impacts of networks on manufacturing systems. Additionally, industry lacks effective and easy-to-use strategies for test and evaluation of such systems in a way that correlates network performance with operational performance. Furthermore, since factory operators desire the ability to control operations within the workcell using wireless technology, we present a novel method to simultaneously capture network and operational event information using a graph database (GDB). The use of a GDB allows for more intuitive inferences to be made through the stored relationships and graph theoretic models [7]. In this paper, we present a GDB approach to the capture and analysis of factory workcell performance utilizing the Neo4j database platform. We present a proposed schema of the database, the process for capturing both network and operational events, and examples of querying the database for cyber-physical performance evaluation of the workcell for our collaborative two-robot machine-tending workcell [8].

II. Graph Databases

A GDB is the type of databases that uses nodes, edges, and properties to store and present data. A GDB is a part of a family of databases known as NoSQL databases that are used often to represent complex interrelated structures of data and their relationships. This can be very difficult with traditional relational databases. The GDB places a high priority on the relationships (edges) between units of information. In addition, the GDB does not enforce any particular schema or structure, and, therefore, provides greater flexibility in storing and representing information in which the parts of information

Candell, Rick; Hany, Mohamed; Liu, Yongkang; Montgomery, Karl; Foufou, Sebti. "A Graph Database Approach to Wireless IIoT Work-cell Performance Evaluation." Paper presented at 2020 IEEE International Conference on Industrial Technology, Buenos Aires, AR. February 26, 2020 - February 28, 2020.

may vary among units. The relationships within a GDB are efficiently queried because they are persistently stored within the database. In a GDB, queries can be made based on relationships. This, in particular, presents an advantage when storing information regarding systems with correlations that are apparent but difficult to visualize or quantify. The GDB approach was chosen for this specific purpose—to quantify and visualize relationships between non-ideal communications within a workcell and its impact on the physical system.

III. Related Work

Multiple surveys about GDBs have been presented to describe the associated models, tools, and their features in [7], [9], [10]. Also, examples of applications and implementations of GDBs are presented in [11] to show their use on enterprise data, social networks, and determining security and access rights. It was found that GDBs provide the much needed structure for storing data and incorporating a dynamic schema. On the other hand, query languages are used to extract data including traversing the database, comparing nodes properties, and subgraph matching [12]. The performance of different GDB tools and methodologies is analyzed and compared in [13], [14]. Multiple comparisons in these articles have shown improved performance of Neo4j in the general features for data storing and querying, and data modeling features such as data structures, query languages and integrity constraints.

Furthermore, industrial data analytics play an essential role in achieving the smart factory vision and improving decision-making in various industrial applications. Five main industrial data methodologies are studied including highly distributed data ingestion, data repository, large-scale data management, data analytics, and data governance [15]. Industrial data processing offers valuable information about various sections of industrial applications including inefficiencies in industrial processes, costly failures and down-times, and effective maintenance decisions [16]. In [17], a platform for performing industrial big data analysis is presented where the performance requirements are introduced to achieve a cost-effective operation. Various other frameworks for industrial data analysis can be found in [18], [19], where the importance of using data analysis in decision making is emphasized.

Due to its advantages including scalability, efficiency, and flexibility, NoSQL databases are a popular alternative to relational databases in the case of large amounts of data in various applications [20]. The GDB is a kind of NoSQL database approaches that additionally handles complex relationships [21]. GDBs are widely adopted in various industry-related applications and use cases such as network operations, fraud detection, and asset and data management [22]. Relationships in social networks have been modeled using a GDB for structural information mining and marketing [23]. On the other hand, business solutions for scenarios with multiple large data sources require distributed processing in decision making for various problems such as fraud detection, trend prediction, and product recommendation [24].

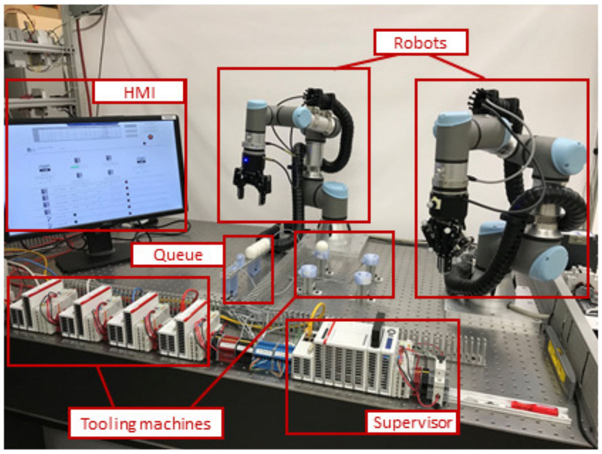


Fig. 1: Collaborative work cell testbed

IV. Case Study: Robotic Machine-tending

In this section, we present a two-robot machine tending workcell case study. We first present the design of the workcell followed by an elaboration of the database design called a schema. The information work-flow, i.e., the process for collecting, processing, and analyzing the network event data is presented. We then provide examples of the resulting graph and results of targeted queries that demonstrate the purpose of the database.

A. Workcell Design

To facilitate wireless network research and showcase the power of wireless technologies in industrial practices, a testbed shown in Fig. 1 has been developed at the National Institute of Standards and Technology (NIST) as described in [8]. The testbed is composed of two collaboration-grade robots, a supervisory programmable logic controller (PLC) used for the control of the workcell, four smaller PLCs serving as computer numerical control (CNC) machine emulators, and a human-machine interface (HMI). Each robot is equipped with a six-degrees-of-freedom (DOF) force-torque (FT) sensor and a two-finger gripper. A Modbus/TCP server is included within the supervisor PLC and is used for communication between the supervisor and the robots. The PLCs themselves communicate to each other using the Beckhoff Automation Device Specification (ADS) protocol. All elements within the workcell are synchronized to a stable and accurate grand-master clock. Therefore, as described in [8], the operational, network, and measurement elements are all synchronized to the grand-master clock through a precision time protocol (PTP)-capable switch.

Work-orders for the workcell are submitted through the supervisor. Each work-order consists of a work-plan for a part, and the work-plan determines how each part moves through the workcell until it is completed. The inspection of each part is conducted at each machining station, and after the final inspection, the part is placed back into the input queue. Under normal operating conditions, the work continues until all work-orders have been processed. This continuous form of operation provides ample opportunity to collect statistically significant metrics of both the network and the operation of the workcell.

B. Database Schema

GDBs are NoSQL databases such that the database does not contain any predefined structure or rules to enforce such structure. This is a major difference between relational databases and GDBs. Nevertheless, it was necessary to sketch a pseudo-schema to capture the intended nodes and relationships that would be stored within the database (the terms pseudo-schema and schema will be used interchangeably). Before describing the schema itself, it is necessary to first explain the requirements of the schema. Therefore, the requirements for our schema are as follows:

Time Any manufacturing automation system is indeed a time-varying control system with network and operational events. The database schema must necessarily support time-based queries and, specifically, time-windowed queries.

Operational Events The schema must represent operational events such as the movement of a robot arm or the movement of a part.

Network Events The schema must represent network events such as the transmission of packets.

Message Grouping The schema must support grouping of logically related events such that those events can be correlated to a specific occurrence within the testbed.

Wireless Support The schema must support the capture of both wired and wireless network traffic without special provisions within queries for either.

QoS Support The schema must allow for the capture of quality of service (QoS) data when available.

Spectrum Monitoring The schema must support the capture and association of network events with observations from a spectrum monitoring system (SMS) if that information is available.

1) Node Design: Given the fore-mentioned requirements, a sample pseudo-schema is shown in Fig. 2, which represents the intended structure of the information within the GDB. It is important to remember that since GDB schemas are not really schemas, such as those found in relational databases, but representations of intent, the schema represented here should be considered a notional example of the final product. Within the graphs, nodes represent logical elements, and edges represent the relationships between those elements. Both nodes and edges may contain properties providing more description and labels that define categories or classes of the said nodes or edges. Our schema is designed such that the data within the graph is intuitive to understand and allows for time-based queries to occur. The facilitation of time-based queries was an essential requirement of our database design. Our schema is represented using the following node labels:

Actor A physical component within the factory workcell such as a robot, PLC, or other networked item.

NtwkID A network address item for an actor such as an Internet Protocol (IP) address.

Transaction A complete information exchange between two or more actors (multiple actors may participate in a transaction).

Message A network transmission event that occurs between two actors (messages are essentially packet transmissions captured at the transport layer; multiple messages support a transaction).

Physical Action A physical occurrence within a factory workcell associated with actors through multiple time based relationships.

SMS An SMS observes and records significant spectral events within the workcell and may report those events to actors within the workcell.

Sniffer Measurement device that records all transmissions conducted over the wireless medium and includes wireless header information for each wireless transmission detected.

Adapter Device that serves to connect an actor to a network (adapters are divided into sub-categories depending on the type of interface to a network).

Adapter:Ethernet A subcategory of adapter representing an Ethernet interface.

Adapter:Wireless A subcategory of adapter representing a wireless interface.

Adapter:Wireless:AP A subcategory of adapter representing a wireless access point interface.

Adapter:Wireless:UE A subcategory of adapter representing a wireless user equipment interface.

QoS Report Quality of service report of a message (not all messages will have a QoS report).

It is important to note that most network infrastructure components are not captured within the graph, but, instead, only basic interfaces between actors and the network are captured. Our intent when designing the graph was to make the graph network and protocol agnostic, such that the network is viewed as a black-box. Accordingly, the captured events of the physical system and the network are considered useful for the analysis of performance.

2) Relationships (Graph Edges): Relationships are edges within the graph that capture the informational interactions between nodes. Relationships, like nodes, can contain labels and properties. As shown in Fig. 2, nodes are connected through defined relationships. A subset of the relationships are defined as follows:

PARTICIPATED_IN Actors will participate in transactions. A transaction exists for each logical set of messages between actors such as the setting of a Modbus register or the sending of a command to a robot. Therefore, actors will participate in many transactions, and multiple actors may participate in a single transaction.

SUPPORTED Messages (i.e., packets between actors) are associated with transactions through the SUPPORTED relationship. Depending on the protocol and the quality

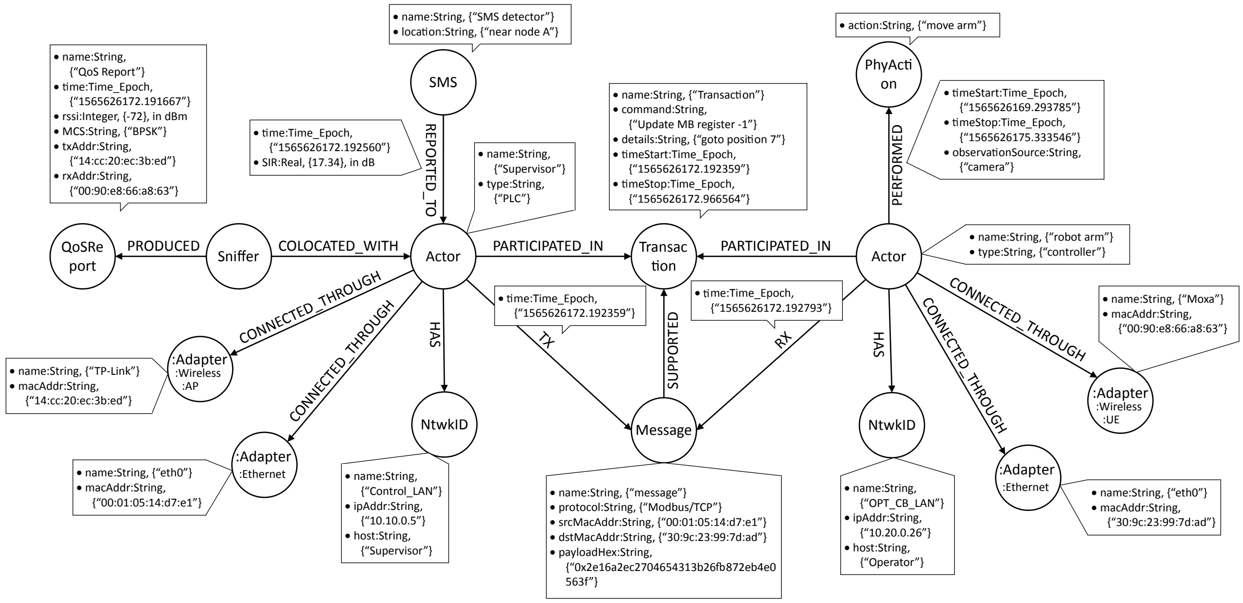

• name:String, {“SMS detector”}
• location:String, {“near node A”}
SMS
REPORTED_TO
• name:String, {“QoS Report”}
• time:Time_Epoch, {“1565626172.191667”}
• rssi:Integer, {-72}, in dBm
• MCS:String, {“BPSK”}
• txAddr:String, {“14:cc:20:ec:3b:ed”}
• rxAddr:String, {“00:90:e8:66:a8:63”}
• time:Time_Epoch, {“1565626172.192560”}
• SIR:Real, {17.34}, in dB
• name:String, {“Supervisor”}
• type:String, {“PLC”}
QoSReport
PRODUCED
Sniffer
COLOCATED_WITH
Actor
PARTICIPATED_IN
• name:String, {“Transaction”}
• command:String, {“Update MB register -1”}
• details:String, {“goto position 7”}
• timeStart:Time_Epoch, {“1565626172.192359”}
• timeStop:Time_Epoch, {“1565626172.966564”}
Transaction
PARTICIPATED_IN
• action:String, {“move arm”}
PhyAction
• timeStart:Time_Epoch, {“1565626169.293785”}
• timeStop:Time_Epoch, {“1565626175.333546”}
• observationSource:String, {“camera”}
PERFORMED
Actor
• name:String, {“robot arm”}
• type:String, {“controller”}
CONNECTED_THROUGH
CONNECTED_THROUGH
HAS
• time:Time_Epoch, {“1565626172.192359”}
TX
SUPPORTED
• time:Time_Epoch, {“1565626172.192793”}
RX
HAS
CONNECTED_THROUGH
CONNECTED_THROUGH
• name:String, {“Moxa”}
• macAddr:String, {“00:90:e8:66:a8:63”}
• name:String, {“TP-Link”}
• macAddr:String, {“14:cc:20:ec:3b:ed”}
:Adapter :Wireless :AP
:Adapter :Wireless :UE
NtwkID
Message
NtwkID
:Adapter :Ethernet
:Adapter :Ethernet
• name:String, {“eth0”}
• macAddr:String, {“00:01:05:14:d7:e1”}
• name:String, {“Control_LAN”}
• ipAddr:String, {“10.10.0.5”}
• host:String, {“Supervisor”}
• name:String, {“message”}
• protocol:String, {“Modbus/TCP”}
• srcMacAddr:String, {“00:01:05:14:d7:e1”}
• dstMacAddr:String, {“30:9c:23:99:7d:ad”}
• payloadHex:String, {“0x2e16a2ec2704654313b26fb872eb4e0563f”}
• name:String, {“OPT_CB_LAN”}
• ipAddr:String, {“10.20.0.26”}
• host:String, {“Operator”}
• name:String, {“eth0”}
• macAddr:String, {“30:9c:23:99:7d:ad”}

Fig. 2: The intended schema (i.e., pseudo-schema) of the graph database used for each operational run of the NIST wireless factory testbed. The graph is organized into nodes and edges, where the edges signify relationships among network elements and physical operational elements.

of the channel, a single transaction could have one or many messages connected through this relationship.

TX/RX An actor may either transmit (TX) or receive (RX) a message. Both the TX and RX relationships contain a timestamp in the format of an epoch time which is a floating point number in seconds since January 1, 1970, with a resolution of microseconds.

PERFORMED When an actor performs a physical action, a relationship is created between the actor and the physical action node. This relationship contains start and stop time properties as well as the source of the observation such as a networked camera.

REPORTED_TO An SMS may be a passive or active listener within a workcell. When an SMS operates as an active listener, spectral reports from the SMS may be sent to an actor such that the actor can respond intelligently to the spectral event. Reports from an SMS to an actor are captured within this relationship.

Other relationships shown in the schema of Fig. 2 but not explained above are considered self-explanatory.

3) Closer Examination: Examining the sample schema more closely, two actor nodes are represented. In this case, Actor A is the supervisory controller, and Actor B is a robot arm. Both nodes participate in a transaction, which, in this example, is a Modbus/TCP exchange. The transaction itself is associated with one or more messages (i.e., packets). Each message associated with a transaction manifests itself as a node in the graph. Multiple message nodes will exist for each transaction. Additionally, QoS reports may be associated with each actor node through a collocated sniffer node. By keeping QoS reports separate, we have the flexibility of supporting different wired and wireless protocols within the same graph. Recall, that a graph database has no enforceable structure and thus affords this type of flexibility. Each actor node may have network identities associated with it through the use of network identifier nodes. Each network identifier node may contain address information such as a hardware address or a network address; however, this is dependent on the protocols and addressing schemes being used, and a node could have many different identification nodes.

4) Physical Actions: Finally, each actor in the graph may associate with physical actions. These actions exist in the database as automatons such that every time a new action occurs, a new edge would be added between the actor and the physical action. Timestamps within the graph represent “measurement time” denoting that all timestamps are accurately synchronized to the grand-master clock. The method of synchronization is outside the scope of this paper and is explained fully in [8]. This paper describes the database structure and the process for preparing and inserting the data into the database, which is described in the following subsection.

C. Information Workflow

The workflow for collecting, processing, and inserting measurement data into the graph database is multi-pronged, as shown in Fig. 3. The workcell, which in our case study is a two-robot workcell with four CNC machines, is instrumented with network probes that capture transmitted and received packets as well as probes to capture operational movements

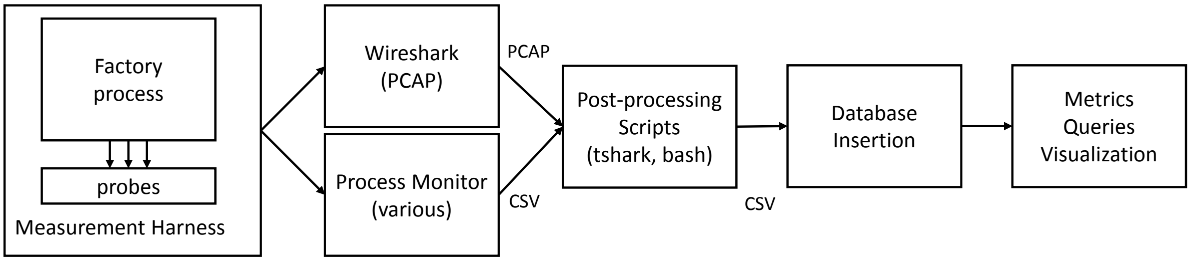

Factory process
probes
Measurement Harness
Wireshark (PCAP)
Process Monitor (various)
PCAP
CSV
Post-processing Scripts (tshark, bash)
CSV
Database Insertion
Metrics Queries Visualization

Fig. 3: Data processing flow from factory workcell to database

such as arm robot position and the state of the supervisor PLC. Network data is stored in packet capture (PCAP) files, while operational data is stored in comma separated value (CSV) files. We developed bash scripts to extract relevant information and prepare the data for insertion into the database. The scripts also contain rules for grouping packets together as transactions based on the application protocol and time. The scripts produce CSV files that are ready for insertion into the Neo4j database. Once the data resides within the database, we apply queries to extract information for the evaluation of workcell performance and visualization of network and operational events within the workcell. By tracking paths through the relationships within the graph, discerning how a network event such as interference is related to physical events such as position uncertainty or part throughput is possible. Various impairments may be introduced as a part of workcell operation. Examples of such impairments include competing wireless traffic, radio interference, and reflections and diffraction due to the multi-path environment [25]. We have shown that it is feasible to implement such impairments and measure the resulting physical performance manifestation [8]. This is accomplished through the use of a radio channel emulator as demonstrated in [26].

D. Sample of a Resulting Graph

In the following, we show results from a trial run of the NIST industrial wireless testbed. In this trial run, a single physical wireless link is used between a wireless adapter connected directly to the supervisor and a wireless access point connected to all the other actors in the testbed. The wireless nodes represent IEEE 802.11b/g/n devices and are connected through a variable radio frequency (RF) attenuator that allows us to vary the channel quality. During the trial run, the production of 10 parts was emulated, which resulted in 10 minutes of network activity.

After populating the database with data captured from the trial run, the resulting realized schema is shown in Fig. 4. The schema visualization is produced by invoking the command

```
call db.schema.visualization()
```

in Neo4j. It is important to note that a realized schema shows only one representation of each node and relationship whereas the intended pseudo-schema shown in Fig. 2 was developed to exemplify the relationships between types of nodes, labels, and relationships. Where label inheritance is employed, such as the case for different adapter types, relationships are reproduced; however, this is a result of the visualization tool rather than the schema itself. Fig. 4 serves, therefore, to validate that the intended schema was indeed realized by the insertion of event data from the testbed. In the realized schema, inherited labels are shown as separate nodes.

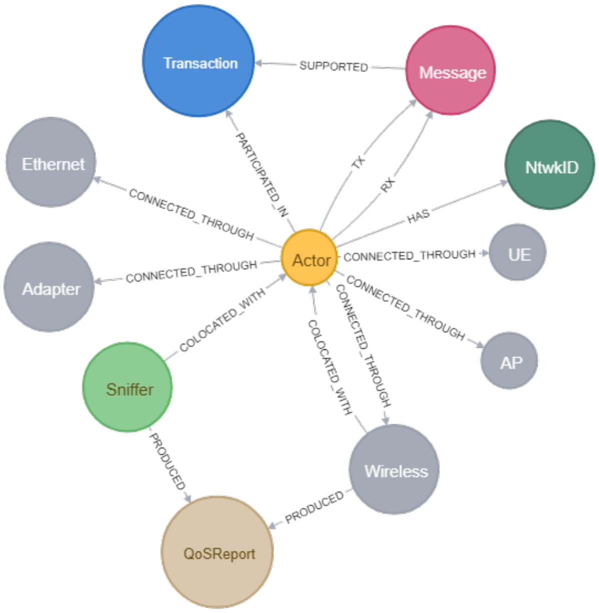


Fig. 4: Realized schema of the graph database fully populated after capturing network and operational data from the NIST industrial wireless testbed.

As described in Section IV-B, the database includes every network transaction that occurs during the operation of the testbed. This includes any logical transaction nodes inserted into the database and any associated packets that happened to traverse the network. Therefore, for a short duration of time depending on packet transmission rates, the amount of data stored in the database can grow quickly. This presents a visualization challenge that graphs are designed to handle. A sample graph is shown in Fig. 5, which represents only 1 second of wired and wireless network data captured from

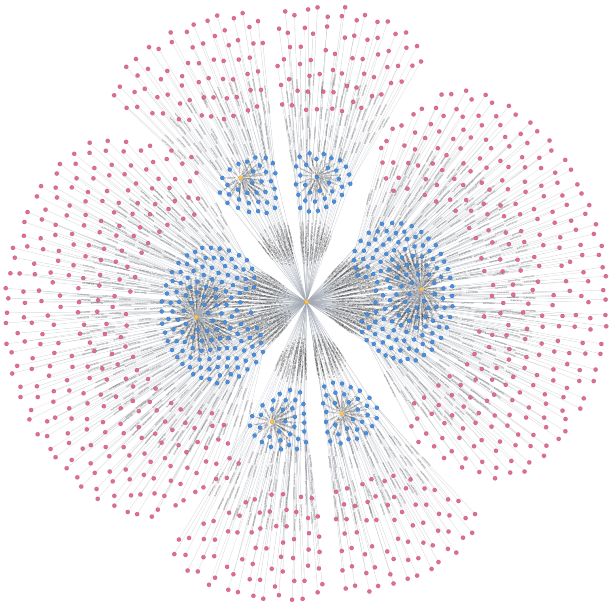


Fig. 5: Visualization of a node graph resulting from a testbed experiment.

the NIST two-robot pick-and-place wireless testbed described in [8].

This visualization is produced by calling the following query in the Neo4j.

```
MATCH p=(a:Actor {name:'Supervisor'})--(t:
    ↪ Transaction)--(b:Actor) ,p2=(m:
    ↪ Message)-->(t) WHERE t.timeStart>T
    ↪ AND t.timeStop<T+1
RETURN p,p2
```

The colors of the resulting nodes follow the realized schema in Fig. 4 while only the actors, transactions, and messages are visualized in 5. The relationships between the messages and transactions are shown where a single transaction is connected to at least two messages to represent the communications between any two actors. The variable T in the query represents an arbitrary time variable in seconds within the trial run to capture the data within 1 second only.

We then show a more detailed visualization for all the nodes and relationships corresponding to a single transaction. This visualization is produced by calling the following query where Ts is an arbitrary time variable to specify the timeStart value for the single transaction.

```
MATCH p=(a:Actor {name:'Supervisor'})-->(t
    ↪ :Transaction {timeStart:Ts})<--(b:
    ↪ Actor {name:"CNC-1"})
WITH p, a, b, t
MATCH p1=(a)-->(m:Message)<--(b), p2=(m)
    ↪ -->(t), p3=(a)-[:HAS]-(),p4=(a)<-[:
    ↪ COLOCATED_WITH]-()-[:PRODUCED]->(q:
    ↪ QoSReport), p5=(a)-[:
    ↪ CONNECTED_THROUGH]-(),p6=(b)-[:HAS
    ↪ ]-(), p7=(b)-[:CONNECTED_THROUGH]-()
WHERE q.time>t.timeStart AND q.time<t.
    ↪ timeStop
RETURN p,p1,p2,p3,p4,p5,p6,p7
```

In Fig. 6, the actor nodes are labeled by their names *CNC-1* and *Supervisor*, the transaction is labeled by its type *ADS*, the messages are labeled by their transmission role *Request* and *Response*, the NtwkIDs are labeled by their IP addresses, the Ethernet adapters are labeled by their names *eth0*, the wireless adapters are labeled by their names *Moxa* and *TP-Link*, the sniffer is labeled by its name *WLS1*, and the QoS reports are labeled by the received signal strength indicator (RSSI) value in dBm. This visualization includes only the QoS reports generated within the duration of the corresponding transaction.

E. Calculation of Metrics

We now present an example of deploying the proposed GDB approach in industrial wireless analysis. During the trial run time, we varied an RF attenuator value in the single wireless link. The attenuation can take the values {0, 10, 20, 30, 40, 50, 60} dB. We evaluate the message latency as the difference between the receive and transmit times of a message including transmission, processing, and retransmission times. Each message has been coupled to a QoS report that is the one reported at the closest time instant before the message transmit time. One of the parameters in the QoS report is the RSSI value captured by the Sniffer colocated with the supervisor's wireless adapter.

In Fig. 7, we present a scatter plot showing the correlation between the message latency in seconds to the measured RSSI values by the sniffer in dBm. The figure shows that the latency is higher at the lower RSSI values due to the increased number of retransmissions. Generally, an IEEE 802.11 transmission can occur using different IEEE 802.11 mode and modulation and coding scheme (MCS) based on the channel quality. At the RSSI of -90 dBm, the receiver should operate at the most robust communications mode and the lowest MCS index. However, retransmissions still occur due to having the received power near the sensitivity of the supervisor's wireless adapter. As shown in Fig. 7, retransmissions may occur at other RSSI values as well when the transmitter selects a higher mode of transmission and a higher MCS index. In this case, we assert that the receiver is operating close to the marginal sensitivity for a given mode (e.g., 802.11n) and MCS index. We observed that the transmitter switches to a lower MCS index for the retransmitted messages, which supports our supposition. Therefore, latency can in fact be high for better RSSI values. This effect is illustrated in Fig. 7 where certain messages can have latency values at -43 dBm, which are comparable to those in the case of -90 dBm.

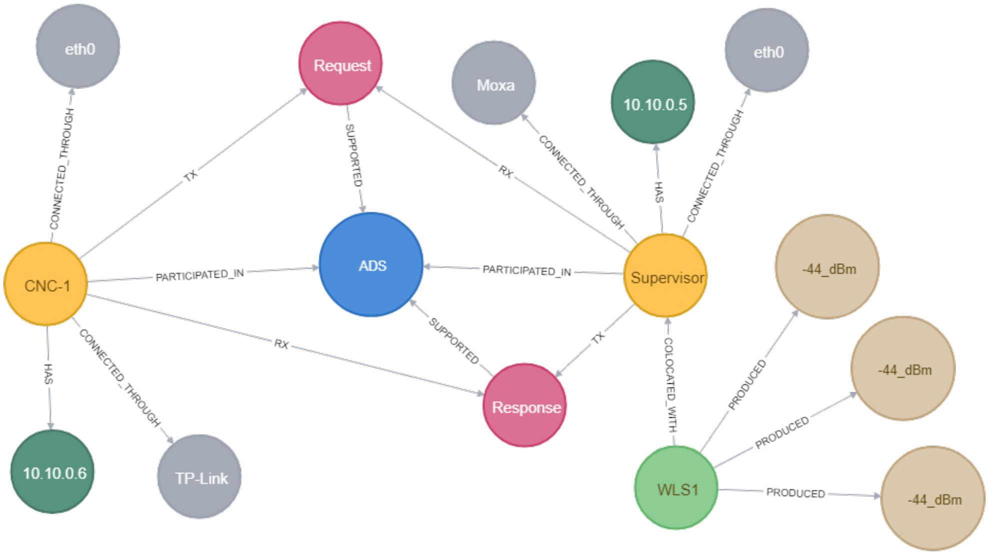


Fig. 6: A detailed visualization resulting from a testbed experiment for all the nodes and relationships corresponding to a single transaction.

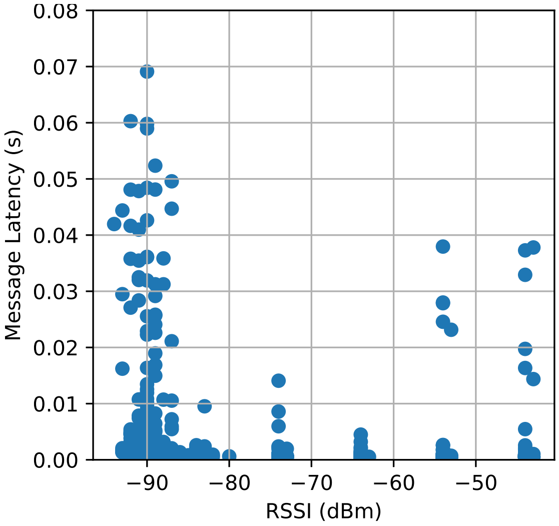


Fig. 7: Correlation between the message latency and the RSSI values by the sniffer.

V. Future Direction

We have presented the general architecture of a performance evaluation database. Future work will include developing query algorithms for the calculation of performance metrics as well as correlation of network events to the physical performance of the manufacturing system. Algorithms will be developed to indicate locations in time of lost information making more direct performance correlations possible. Beyond performance evaluation of the manufacturing system, we believe that the graph database approach enables anomaly detection because relationships within the data are intrinsically stored and thus efficiently queried. The structure of the information is important within a graph database, and defining the relationships such that the paths through nodes may be discovered is essential to efficient queries and discovery of hidden correlations. Therefore, more research will be done in the modification of the presented schema. Machine learning will be applied for the detection of anomalies, performance degradation, signal quality, and network performance enhancement through more efficient resource allocation. Our plans also include online database insertions to enable the demonstration of online operation of the database. We also intend to examine the development of better control system strategies that mitigate network losses and delays. Since this approach is indifferent to the communications protocols used, we intend to extend our database approach to include various protocols, radio bands such as millimeter wave bands for 5G communications, and different use cases to include mobile robotic platforms and safety critical systems.

VI. Conclusions

We have presented in this paper a novel approach to capturing network and operational event information from a factory workcell with the purposes of 1) capturing and storing of network and operational events, 2) calculating performance metrics of the network, and 3) discovering performance dependencies between the network and the physical assembly of the workcell. Using a graph database, we have demonstrated that it is possible to construct such a database, compute network performance metrics and discover correlations. We have also developed the capability of examining the correlation between network events and the performance of physical actions. This will be a source of further research.

Future progress and measurement data will be deposited in the NIST public domain repository as a reference for industrial traffic modeling efforts and comparative studies on industrial wireless technologies [27].

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] L. L. Bello, J. Akerberg, M. Gidlund, and E. Uhlemann, "Guest editorial special section on new perspectives on wireless communications in automation: From industrial monitoring and control to cyber-physical systems," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1393–1397, June 2017.

[2] V. K. L. Huang, Z. Pang, C. A. Chen, and K. F. Tsang, "New trends in the practical deployment of industrial wireless: From noncritical to critical use cases," *IEEE Industrial Electronics Magazine*, vol. 12, no. 2, pp. 50–58, June 2018.

[3] (ETSI), "Etsi tr 103 588 reconfigurable radio systems (rrs); feasibility study on temporary spectrum access for local high-quality wireless networks," Sophia-Antipolis, 2018.

[4] H. Kagermann, W. Wahlster, and J. Helbig, "Recommendations for implementing the strategic initiative industrie 4.0, industrie 4.0 working group," 2013.

[5] A. Barnard Feeney, S. Frechette, and V. Srinivasan, *Cyber-Physical Systems Engineering for Manufacturing*. Cham: Springer International Publishing, 2017, pp. 81–110. [Online]. Available: https://doi.org/10.1007/978-3-319-42559-74_4

[6] Wireless Broadband Alliance (WBA), "Enhanced Wi-Fi - 802.11ax Decoded: Overview, Features, Use Cases and 5G Context." [Online]. Available: https://www.wballiance.com/wp-content/uploads/2018/09/Enhanced-802.11ax-Overview-Use-Cases-Features-5G-Context-v1.asd-Final-Clean.pdf

[7] R. Angles and C. Gutierrez, "Survey of graph database models," *ACM Comput. Surv.*, vol. 40, no. 1, pp. 1:1–1:39, Feb. 2008. [Online]. Available: http://doi.acm.org/10.1145/1322432.1322433

[8] Y. Liu, R. Candel, M. Kashef, and K. Montgomery, "A collaborative work cell testbed for industrial wireless communications the baseline design," in *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, June 2019, pp. 1315–1321.

[9] R. Kumar Kaliyar, "Graph databases: A survey," in *International Conference on Computing, Communication Automation*, May 2015, pp. 785–790.

[10] H. R. Vyawahare and P. P. Karde, "An overview on graph database model," *International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)*, vol. 3, pp. 7454–7457, August 2015.

[11] K. N. Satone, "Modern graph databases models," *International Journal of Engineering Research and Applications (IJERA)*, pp. 19–24, 2014.

[12] P. T. Wood, "Query languages for graph databases," *SIGMOD Record*, vol. 41, pp. 50–60, 2012.

[13] P. S. Jadhav and R. K. Oberoi, "Comparative analysis of graph database models using classification and clustering by using weka tool," 2015.

[14] P. Macko, D. Margo, and M. Seltzer, "Performance introspection of graph databases," in *Proceedings of the 6th International Systems and Storage Conference*, ser. SYSTOR '13. New York, NY, USA: ACM, 2013, pp. 18:1–18:10. [Online]. Available: http://doi.acm.org/10.1145/2485732.2485750

[15] J. Wang, W. Zhang, Y. Shi, S. Duan, and J. Liu, "Industrial big data analytics: Challenges, methodologies, and applications," *CoRR*, vol. abs/1807.01016, 2018. [Online]. Available: http://arxiv.org/abs/1807.01016

[16] J. Lee, *Industrial Big Data (Mechanical Industry Press, China)*, 07 2015.

[17] (GE), "Unlocking machine data to turn insights into powerful outcomes." [Online]. Available: https://www.ge.com/digital/

[18] Brian Courtney, "Industrial big data analytics: The present and future." [Online]. Available: https://www.isa.org/intech/20140801/

[19] (ABB), "Big Data and decision-making in industrial plants." [Online]. Available: https://new.abb.com/cpm/production-optimization/big-data-analytics-decision-making

[20] J. Pokorny, "NoSQL databases: a step to database scalability in web environment," *International Journal of Web Information Systems*, vol. 9, no. 1, pp. 69–82, 2013. [Online]. Available: https://doi.org/10.1108/17440081311316398

[21] Z. J. Zhang, "Graph databases for knowledge management," *IT Professional*, vol. 19, no. 6, pp. 26–32, November 2017.

[22] J. Webber and I. Robinson, "The top 5 use cases of graph databases," *white paper, Neo4j*, 2015.

[23] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Discov. Data*, vol. 5, no. 4, pp. 21:1–21:37, Feb. 2012. [Online]. Available: http://doi.acm.org/10.1145/2086737.2086741

[24] S. Skhiri and S. Jouili, *Large Graph Mining: Recent Developments, Challenges and Potential Solutions*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 103–124. [Online]. Available: https://doi.org/10.1007/978-3-642-36318-4_5

[25] R. Candell, C. Remley, J. Quimby, D. Novotny, A. Curtin, P. Papazian, G. Koepke, J. Diener, and M. Kashef, "Industrial wireless systems: Radio propagation measurements," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 2017. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.1951.pdf

[26] R. Candell, K. Montgomery, M. Kashef, Y. Liu, and S. Foufou, "Wireless interference estimation using machine learning in a robotic force-seeking scenario," in *2019 IEEE 28th International Symposium on Industrial Electronics (ISIE)*, June 2019, pp. 1334–1341.

[27] R. Candell, M. Hany, Y. Liu, and K. Montgomery, "Reliable, High Performance Wireless Systems for Factory Automation," 2019. [Online]. Available: https://www.nist.gov/programs-projects/reliable-high-performance-wireless-systems-factory-automation

Development and Applications of a Four-Volt Josephson Arbitrary Waveform Synthesizer

N. E. Flowers-Jacobs, A. Rüfenacht, A. E. Fox, S. B. Waltman, R. E. Schwall, J. A. Brevik, P. D. Dresselhaus, and S. P. Benz
National Institute of Standards and Technology (NIST), Boulder, CO 80305 USA
nathan.flowers-jacobs@nist.gov

***Abstract*— We have recently created a 4 V rms cryocooled JAWS (Josephson Arbitrary Waveform Synthesizer) using 204,960 nearly identical Josephson junctions (JJs) that are embedded in coplanar-wave guides. The JJs are pulse-biased at repetition rates up to 16×10^9 pulses per second to create quantum-accurate, calculable AC waveforms at frequencies from DC to greater than 1 MHz. This system has metrological applications including in precision ac voltage calibrations, comparisons of arbitrary impedances, and ac power measurements.**

Keywords*— *Metrology, digital-to-analog conversion, Josephson junction arrays, measurement standards, signal synthesis, superconducting integrated circuits, voltage measurement.

I. Introduction

Over the past thirty years, electrical metrology has been revolutionized by superconducting Josephson-junction (JJ) circuits' ability to generate quantum-accurate voltages. Circuits and systems with higher output voltage have enabled these advances in metrology. We recently doubled the output voltage of the JAWS (Josephson Arbitrary Waveform Synthesizer) system compared to previous systems [1] by developing a JAWS chip that generates dc and ac voltages up to 1 MHz with rms amplitudes up to 2 V. This performance improvement has enabled the direct calibration of new instrument ranges and has opened new applications in impedance and ac power metrology.

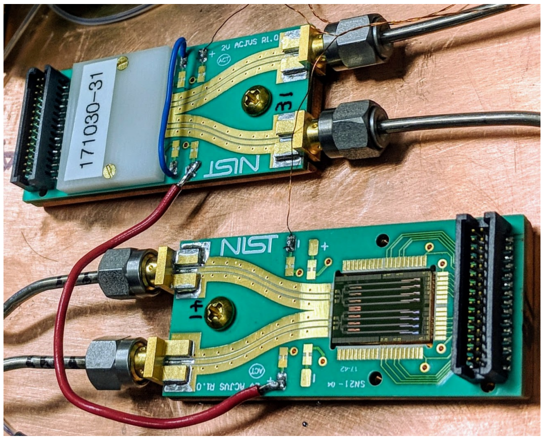


Fig. 2. Picture of two 2 V rms JAWS chips on a 4 K cryocooler

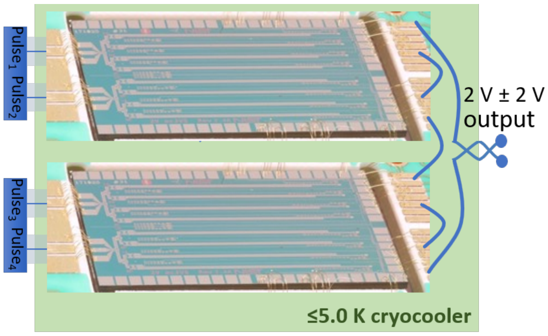


Fig. 1. Simplified system diagram of a 4 V rms cryocooled JAWS

II. Recent Progress

We created a 4 V rms cryocooled system (see Figure 1) by using two of the 2 V JAWS chips containing a combined total of 204,960 nearly identical JJs [2]. The JJs are arranged in 16 series arrays that are embedded in separate coplanar-wave guides (see Figure 2). The quantum-accurate JAWS waveforms are based on a delta-sigma encoding of the JJ voltage pulses which have an integrated area *h/2e*. The voltage pulse timing is controlled by current biasing the JJs with a programmable series of fast current pulses at up to 16×10^9 pulses per second. Each 2 V chip has two high-speed pulse inputs, each of which use two layers of superconducting Wilkinson dividers to split their respective inputs into four signals that each bias one series JJ array. To generate a quantum-accurate output voltage, each input bias pulse must induce every JJ to create a single quantized output pulse.

The 4 V rms output voltage of the new JAWS system has allowed us to calibrate precision ac instruments on higher voltage ranges. Other applications require the ability of these 2 V JAWS chips to generate two separate independent arbitrary waveforms that each have an amplitude up to 1 V rms. For example, we have used these dual 1 V quantum-accurate signals for impedance metrology. The two independently programable JAWS signals are combined with a digital bridge to replace the transformer-based sources typically used in precision impedance bridges. The new JAWS-digital bridge impedance systems have simplified and improved the uncertainty of comparisons between arbitrary impedances [3]. Another new application for JAWS circuits is ac power metrology, where dual quantum-accurate programmable signals with defined harmonic content having specified amplitudes and phases will provide direct quantum-based reference signals for calibrating electric power meters [4].

Acknowledgment

The authors would like to thank M. M. Elsbury and Z. Popović for developing microwave designs, and the Boulder MicroFabrication Facility for facilitating chip fabrication.

References

[1] N. E. Flowers-Jacobs, A. E. Fox, P. D. Dresselhaus, R. E. Schwall, and S. P. Benz, “Two-Volt Josephson Arbitrary Waveform Synthesizer Using Wilkinson Dividers,” *IEEE Trans. Appl. Supercond.*, vol. 26, no. 6, pp. 1–7, Sep. 2016.

[2] A. E. Fox, P. D. Dresselhaus, A. Rüfenacht, A. Sanders, and S. P. Benz, “Junction yield analysis for 10 V programmable Josephson voltage standard devices,” IEEE Trans. Appl. Supercond., vol. 25, no. 3, Jun. 2015.

[3] F. Overney, N. E. Flowers-Jacobs, B. Jeanneret, A. Rüfenacht, A. E. Fox, J. M. Underwood, A. D. Koffman, and S. P. Benz, “Josephson-based full digital bridge for high-accuracy impedance comparisons,” *Metrologia*, vol. 53, no. 4, pp. 1045–1053, Aug. 2016.

[4] B. C. Waltrip, B. Gong, Y. Wang, C. J. Burroughs, A. Rüfenacht, S. P. Benz, P. D. Dresselhaus, “AC Power Standard Using a Programmable Josephson Voltage Standard,” *IEEE Trans. Instrum. Meas.*, vol. 58, no. 4, pp. 1041–1048, Apr. 2009.

Flowers-Jacobs, Nathan; Rufenacht, Alain; Fox, Anna; Waltman, Steven; Schwall, Robert; Brevik, Justus; Dresselhaus, Paul; Benz, Samuel P.
”Development and Applications of a Four-Volt Josephson Arbitrary Waveform Synthesizer.” Paper presented at 2019 IEEE International Superconductive Electronics Conference (ISEC), Riverside, CA, US. July 28, 2019 - August 01, 2019.

Sep. 14, 2019 (Final Rev) For: *Proceedings of COMSOL Users Conference, Boston, Oct. 2-4. 2019*
https://www.comsol.com/conference2019/view-paper-file/xxxxx

Design of an intelligent PYTHON code to run coupled and license-free finite-element and statistical analysis software for calibration of near-field scanning microwave microscopes[1]

Jeffrey T. Fong[2*], N. Alan Heckert[3], James J. Filliben[3], Pedro V. Marcal[4], Samuel Berweger[5], T. Mitchell Wallis[5], Kristen Genter[5], and Pavel Kabos[5]
1. Contribution of the National Institute of Standards & Technology (NIST). Not subject to copyright.
2. Applied & Computational Mathematics Division, NIST, Gaithersburg, MD 20899-8910, U.S.A.
*Corresponding author contact, *fong@nist.gov*, or, *fong70777@gmail.com*
3. Statistical Engineering Division, NIST, Gaithersburg, MD 20899-8960, U.S.A.
4. MPact Corp., Oak Park, CA 91377, U.S.A.
5. Applied Physics Division, NIST, Boulder, CO 80301, U.S.A.

Abstract: To calibrate near-field scanning microwave microscopes (NSMM) for defect detection and characterization in semiconductors, it is common to develop a parametric finite element analysis (FEA) code to guide the microscope user on how to optimize the settings of the instrument to improve its performance. Two problems arise that make the application of the FEA code difficult if not impossible. The first problem is due to the approximate nature of the FEA method and the critical requirement that the accuracy of the FEA solutions be mathematically verified during the entire calibration process. The second problem is a prerequisite that the user's computer be licensed with the specific FEA software at a sizable cost and training time to the user. In this paper, we solve both problems by designing an intelligent PYTHON code that manages the seamless running of two license-free codes, namely, a compiled parametric COMSOL AC/DC-Module-based code that yields a series of solutions at various finite element mesh densities as input to a FEA-verification code written in a statistical analysis software named DATAPLOT that uses a nonlinear least squares method to check and verify the FEA solution of the COMSOL code. An example of a generic NSMM calibration code running a coupled and license-free finite element and statistical analysis software is presented and discussed.

Keywords: Computational modeling, COMSOL, DATAPLOT, electromagnetics, element type, FEM, finite element method, logistic function, mesh density, near-field scanning microwave microscopy, nonlinear least squares method, PYTHON, statistical analysis, uncertainty quantification.

Disclaimer: Certain commercial equipment, materials, or software are identified in this paper in order to specify the computational procedure adequately. Such identification is not intended to imply endorsement by NIST, nor to imply that the equipment, materials, or software identified are necessarily the best available for the purpose.

1. Introduction

To calibrate near-field scanning microwave microscopes (NSMM), as shown schematically in Fig. 1, for detection, sizing, and depth estimation of surface and subsurface defects in semiconductors, it is common to develop a parametric finite element

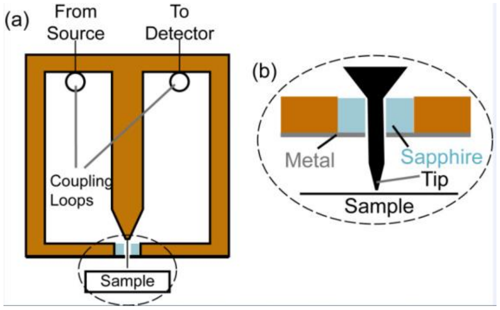


Figure 1. A near-field scanning microwave microscope (NSMM) design based on a coaxial resonator. (a) A schematic of the coaxial resonator, which is operated in transmission mode. (b) A closer view of the prober tip, which is connected to the center conductor of the resonator and extends toward the sample through an opening at the bottom of the resonator (after Gao and Xiang [2] as reproduced in Fig. 7.6, p.115 of a book by Wallis and Kabos [1]).

Fong, Jeffrey T.; Heckert, N. Alan; Filliben, James J.; Marcal, Pedro; Berweger, Samuel; Wallis, Thomas Mitchell (Mitch); Kabos, Pavel. "Design of an intelligent PYTHON code to run coupled and license-free finite-element and statistical analysis software for calibration of near-field scanning microwave microscopes." Paper presented at COMSOL Users Conference, Boston, MA, US. October 02, 2019 - October 04, 2019.

analysis (FEA) code to guide the microscope user on how to optimize the settings of the instrument to improve its performance.

Two problems arise that make the application of the FEA methodology difficult if not impossible. The first problem is due to the approximate nature of the FEA method and the critical requirement that the accuracy of the FEA solutions be mathematically verified during the entire calibration process (see, e.g., refs. [3] through [7]). The second problem is a pre-requisite that the user's computer be licensed with a specific FEA software capable of solving the calibration problem (such as COMSOL [8]) at a sizable cost and training time to the user.

In this paper, we solve both problems by formulating the conceptual design of an intelligent PYTHON code that manages, as shown by Chollet [9] and Fong, et al [10], the seamless running of two license-free[6] codes, namely, a compiled [11] parametric COMSOL AC/DC-Module-based code that yields a series of solutions at various finite element mesh densities as input to a FEA-verification code written in a statistical analysis software named DATAPLOT [12] that uses a nonlinear least squares logistic function fit method [6, 7] to verify the FEA solution of the COMSOL code.

To present the design concept of an intelligent PYTHON code for NSMM calibration, we choose to use an example problem with a very simple sample geometry, namely, a small spherical inclusion in a semiconductor block of 2 um length, 1 um width, and 0.5 um thickness. The diameter of the inclusion is 25 nm, and the clear distance from the top of the sphere to the top surface of the block is 33.3 nm.

In addition, we prescribe that the electrode radius is 10 nm, the electrode voltage is 0.1 v, and the operating frequency is 5 GHz.

To simplify the problem further, we also prescribe the material properties of the block and the inclusion without an estimate of uncertainty, namely,

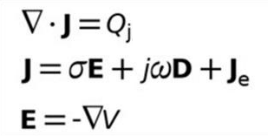
The block electric conductivity = 1e-12 S/m.
The block dielectric constant = 11.7.
The inclusion electric conductivity = 0.01 S/m.
The inclusion relative permittivity = 20.

In Section 2, we develop a parametric finite element analysis (FEA) code using the AC/DC module of the COMSOL code [8] and the COMSOL Compiler code [11] to obtain an executable code that runs in any Windows-based laptop without a COMSOL license.

In Section 3, we develop a FEA verification code written in DATAPLOT [12] to check whether the finite element solutions at increasing mesh densities converge to an asymptotic solution with an acceptable measure of uncertainty. In Section 4, we design a PYTHON code to manage the running of a coupled FEA code and a verification code such that a verified FEA solution is achieved to a prescribed level of uncertainty as required by the calibration mission.

2. Finite Element Analysis Code for a Simple NSMM Flaw Depth/Size Estimation Problem

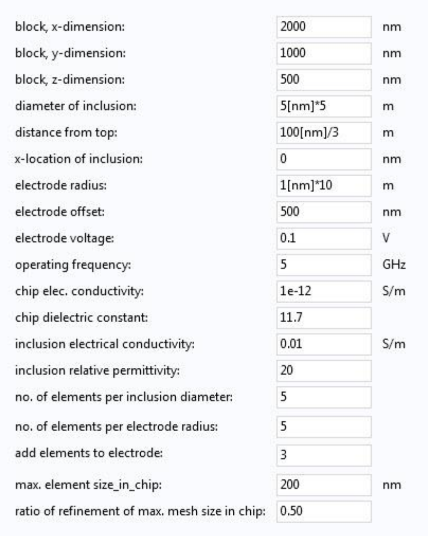
In Figs. 2 and 3, we show the governing equations and a list of 19 adjustable parameters for the COMSOL FEA code of the simple NSMM problem, respectively. In Figs. 4, 5, and 6, we show



$$\nabla \cdot \mathbf{J} = Q_j$$
$$\mathbf{J} = \sigma\mathbf{E} + j\omega\mathbf{D} + \mathbf{J}_e$$
$$\mathbf{E} = -\nabla V$$

Note: **J** is the current density, Q_j, the current source, σ, the conductivity, **E**, the electric field, ω, the frequency, **D**, the electric flux density, $\mathbf{J_e}$, the external current density, *V*, the electric potential, and $j = \sqrt{-1}$.

Figure 2. Governing equations of the example problem as shown in the COMSOL AC/DC Module [8] screen output.



block, x-dimension:	2000	nm
block, y-dimension:	1000	nm
block, z-dimension:	500	nm
diameter of inclusion:	5[nm]*5	m
distance from top:	100[nm]/3	m
x-location of inclusion:	0	nm
electrode radius:	1[nm]*10	m
electrode offset:	500	nm
electrode voltage:	0.1	V
operating frequency:	5	GHz
chip elec. conductivity:	1e-12	S/m
chip dielectric constant:	11.7	
inclusion electrical conductivity:	0.01	S/m
inclusion relative permittivity:	20	
no. of elements per inclusion diameter:	5	
no. of elements per electrode radius:	5	
add elements to electrode:	3	
max. element size_in_chip:	200	nm
ratio of refinement of max. mesh size in chip:	0.50	

Figure 3. A list of parameters in the COMSOL FEA code that are designed to be changed by a user as needed.

[6] COMSOL's license terms state that a code generated by the COMSOL Compiler may be executed by anyone without the need to access a license file. Thus, the generated program can be run by anyone, including those who do not have a paid COMSOL subscription. In this paper we use the term "license-free" when describing these particular terms-of-use for a COMSOL Compiler-generated code**.** It does not apply to DATAPLOT, a non-commercial code.

Fong, Jeffrey T.; Heckert, N. Alan; Filliben, James J.; Marcal, Pedro; Berweger, Samuel; Wallis, Thomas Mitchell (Mitch); Kabos, Pavel. "Design of an intelligent PYTHON code to run coupled and license-free finite-element and statistical analysis software for calibration of near-field scanning microwave microscopes." Paper presented at COMSOL Users Conference, Boston, MA, US. October 02, 2019 - October 04, 2019.

the typical sample geometry and mesh design. After we completed the COMSOL code, we compiled it to get an executable code that runs in a license-free laptop (see Fig. 7 for a screen output display of the executable code). Typical results of either the original or the compiled code appear in Figs. 8 and 9.

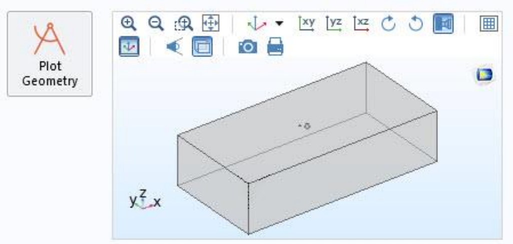


Figure 4. A screen display of the semiconductor block with a small spherical inclusion buried below the top.

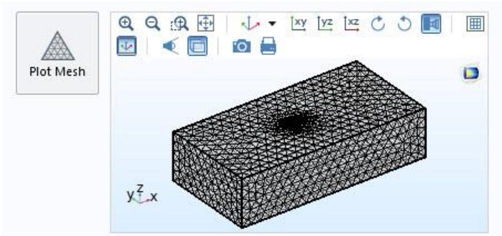


Figure 5. A typical finite element mesh design for the semiconductor block with a buried spherical inclusion where all the elements are tetrahedrons. It is worth noting that meshing in finite element analysis of electromagnetism problems differs from that of structural problems in one important feature, namely, the degrees of freedom of electromagnetic elements are associated with edges, faces, cells, not nodes. For brevity, we shall make no attempt in this paper to explain in details how the meshing was done in a COMSOL code except by referring to its manual [8].

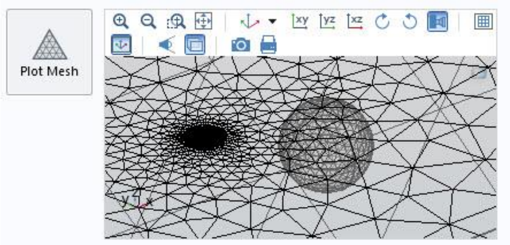
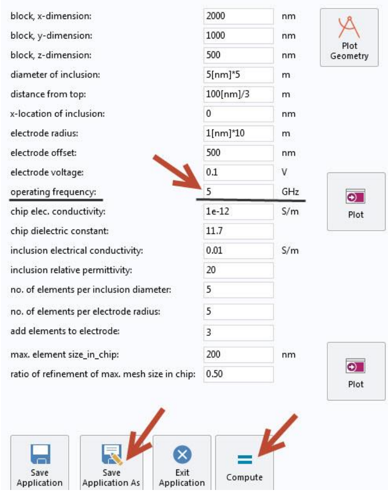


Figure 6. An enlarged view of the finite element mesh design for the buried spherical inclusion with a simultaneous display of the surrounding mesh without the presence of the inclusion. For details, see manual [8].



block, x-dimension:	2000	nm
block, y-dimension:	1000	nm
block, z-dimension:	500	nm
diameter of inclusion:	5[nm]*5	m
distance from top:	100[nm]/3	m
x-location of inclusion:	0	nm
electrode radius:	1[nm]*10	m
electrode offset:	500	nm
electrode voltage:	0.1	V
operating frequency:	5	GHz
chip elec. conductivity:	1e-12	S/m
chip dielectric constant:	11.7	
inclusion electrical conductivity:	0.01	S/m
inclusion relative permittivity:	20	
no. of elements per inclusion diameter:	5	
no. of elements per electrode radius:	5	
add elements to electrode:	3	
max. element size_in_chip:	200	nm
ratio of refinement of max. mesh size in chip:	0.50	

Figure 7. Screen output display of running a compiled COMSOL FEA code after the original COMSOL code is compiled. Note the presence of all adjustable parameters and buttons to re-compute and to save results of a new run.

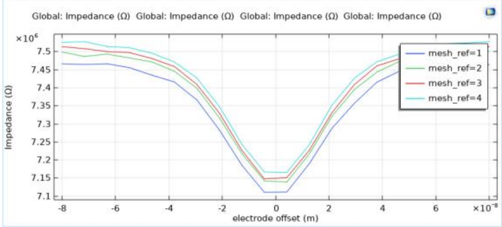


Figure 8. Impedance results of the first four runs at four different finite element mesh densities to check convergence of solutions.

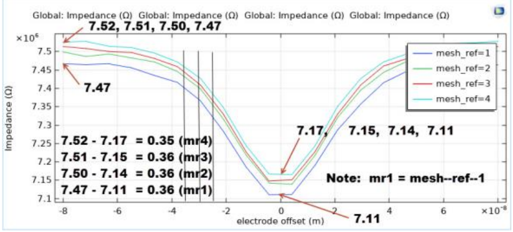


Figure 9. Readings for a key result of the FEA runs for four different mesh densities. When the probe is located directly over the center of the spherical inclusion, the values of the impedance correspond to four mesh densities are:

Mesh Ref. No.	d.o.f.	Impedance (million ohms)
1 (coarser)	51,020	7.11
2	52,058	7.14
3	53,537	7.15
4 (finer)	55,313	7.17

3. A Finite Element Solution Verification Code

To obtain an asymptotic solution of a sequence of FEA candidate solutions at increasing mesh densities (or degrees of freedom, d.o.f.), we apply a 4-parameter nonlinear least-squares (NLLSQ) logistic function fit method (see, e.g., Fong, et al. [6, 7]). A minimum of five candidate solutions is required to run the fit method. For the results given in Fig. 9, we had only four candidate solutions, so we need to run the executable FEA code one more time to generate a fifth candidate solution as follows:

Mesh Ref. No.	d.o.f.	Impedance (million ohms)
1 (coarser)	51,020	7.11
2	52,058	7.14
3	53,537	7.15
4 (finer)	55,313	7.17
5 (new)	**57,089**	**7.18**

In Fig. 10, we show a plot of the NLLSQ fit of 5 points with an asymptotic solution = 7.186 M-ohms. In Fig. 11, we re-plot the result of the 5-point fit with an estimate of uncertainty defined at one million d.o.f. For the calibration purpose in mind, that uncertainty (= 14 %) is too large, so we need to add more points until the uncertainty is less than 2 %. In Figs. 12-14, we show that the goal is reached where the number of candidate solutions reaches nine.

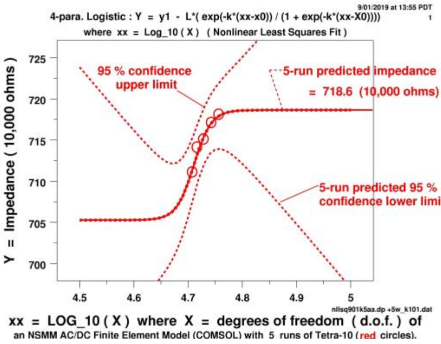


Figure 10. A NLLSQ logistic function fit of 5 candidate solutions yields an asymptotic solution = 7.186 M-ohms**.**

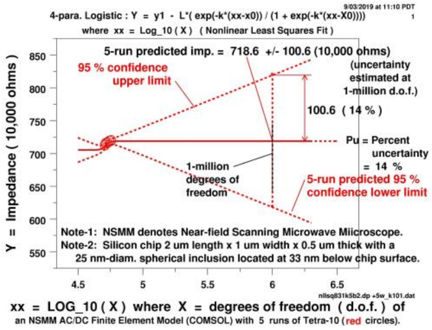


Figure 11. A re-plot of the NLLSQ fit of the 5 candidate solutions with 95 % confidence limits and an estimate of the uncertainty as defined for a million d.o.f. The asymptotic solution now equals 7.186 +/- 1.006 M-ohms.

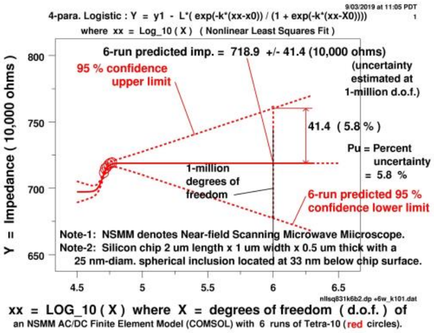


Figure 12. A NLLSQ fit of 6 candidate solutions yields an asymptotic solution equal to 7.189 +/- 0.414 M-ohms.

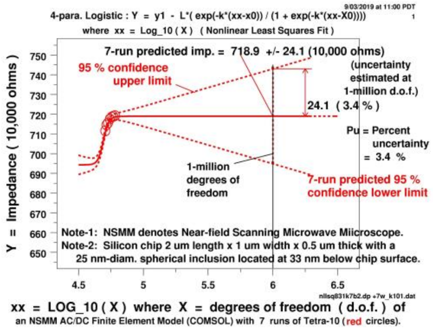


Figure 13. A NLLSQ fit of 7 candidate solutions yields an asymptotic solution equal to 7.189 +/- 0.241 M-ohms.

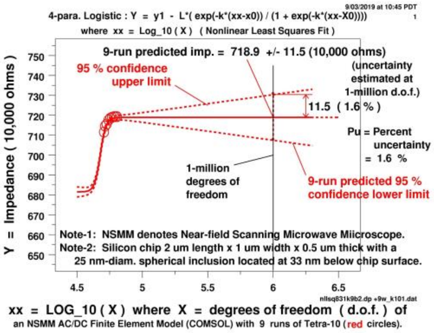


Figure 14. A NLLSQ fit of 9 candidate solutions yields an asymptotic solution equal to 7.189 +/- 0.115 M-ohms.

4. Design of a PYTHON Calibration Code

As shown by Chollet [9], PYTHON is highly suited as a language for writing AI codes in general, because PYTHON code acts as a manager to call on all types of application codes to run on different platforms either sequentially, iteratively, or both.

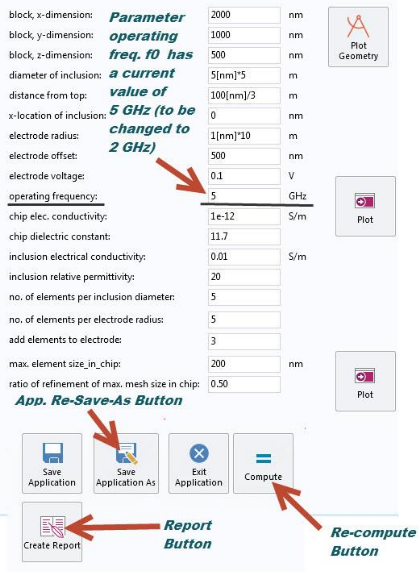
For example, the following chunk of codes, written in PYTHON, is part of a larger code that will allow a user to run a COMSOL application code named "Fong_Kabos_app_Altasim_paid_license.exe":

```
Def InverseProgram() :
  self_m_fileInput=True
  print '***  Run ComsolManager.py '
  total_volume=0.0
  if self_m_fileInput :
    returncode=None
    try :
      z=y[0]+'.dat'
      comsol_data=os.path.join(dir_name,z)
      #returncode= call(['C:/
Fong_Kabos_app_Altasim_paid_license.exe' ,
comsol_data])
      returncode= subprocess.Popen(['C:/
Fong_Kabos_app_Altasim_paid_license.exe' ,
comsol_data])
      #subprocess.Popen.terminate()
      pass
      if returncode :
        print 'Failure with return code' ,returncode
    except :
      pass
InverseProgram()
```

In short, a PYTHON code can be written to run two application codes, namely, a parametric FEA code to solve an NSMM problem and a FEA verification code to obtain an asymptotic solution within a prescribed level of uncertainty. Since it is desirable for all NSMM calibration codes to run on Windows-based computers without FEA software licenses, we show in this paper that with a compiled COMSOL finite element code and an open-source DATAPLOT statistical analysis code, it is feasible to design an intelligent PYTHON code to calibrate instruments such as the near-field scanning microwave microscope.

5. A License-free Calibration Code

The availability of a compiler for COMSOL FEA codes makes it possible for engineers to design all kinds of parametric FEA codes that will run in any Windows laptop without a FEA software license. In Fig. 15, we show how to change the operating frequency from 5 GHz to 2 GHz and re-compute to get new results in a laptop without a software license. The new results are given in Fig. 16. Using the PYTHON code, we can obtain a FEA asymptotic solution to any prescribed degree of uncertainty.



block, x-dimension:	2000	nm
block, y-dimension:	1000	nm
block, z-dimension:	500	nm
diameter of inclusion:	5[nm]*5	m
distance from top:	100[nm]/3	m
x-location of inclusion:	0	nm
electrode radius:	1[nm]*10	m
electrode offset:	500	nm
electrode voltage:	0.1	V
operating frequency:	5	GHz
chip elec. conductivity:	1e-12	S/m
chip dielectric constant:	11.7	
inclusion electrical conductivity:	0.01	S/m
inclusion relative permittivity:	20	
no. of elements per inclusion diameter:	5	
no. of elements per electrode radius:	5	
add elements to electrode:	3	
max. element size_in_chip:	200	nm
ratio of refinement of max. mesh size in chip:	0.50	

Parameter operating freq. f0 has a current value of 5 GHz (to be changed to 2 GHz)

App. Re-Save-As Button

Report Button

Re-compute Button

Figure 15. A screen display of the compiled COMSOL code with 19 adjustable parameters and buttons to re-compute, save new application, and create report.

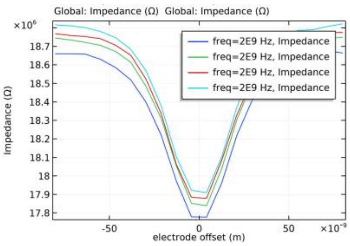


Figure 16. Impedance results for the simple NSMM spherical inclusion problem where the operating frequency has just been altered from 5 GHz to 2 GHz. Since the run only produced 4 candidate solutions, and the NLLSQ fit method requires a minimum of 5 candidates, the intelligent PYTHON code will automatically activate the executable COMSOL code to produce a 5th point.

6. Significance and Limitations of the Conceptual Design of an Intelligent Calibration Code

The conceptual design of an intelligent PYTHON code to run license-free coupled FEA and verification codes is significant in at least two innovative aspects of engineering, namely, (a) it removes a dependency on FEA license for heavy use of candidate runs to achieve solution accuracy, and (b) it provides the engineer access to a sophisticated FEA solution verification tool that would otherwise be unused for lack of lengthy training. It is obvious that the idea of designing an intelligent calibration code for NSMM is not limited only to a specimen with a spherical inclusion, because the finite element method is quite general. However, the proposed design is not without limitations, chiefly among which is the limited emphasis on seeking an asymptotic solution based on mesh density variations. It ignores uncertainty due to other factors such as material properties, voltage and frequency accuracy.

7. Concluding Remarks

Finite element analysis (FEA) methodology is a powerful tool to assist engineers and scientists in calibrating precision instruments such as near-field scanning microwave microscopes (NSMM). The availability of a compiler for any FEA software to convert the source code into an executable code that runs in a laptop without a software license is a major step forward to broaden the usage of FEA methodology. The introduction of an intelligent PYTHON code to run seamlessly both the FEA executable code and a FEA verification code in order to achieve a verified FEA asymptotic solution to a prescribed degree of uncertainty is a much-needed addition to the calibration group of the scientific measurement community.

8. References

1. Wallis, T. M., and Kabos, P., *Measurement Techniques for Radio Frequency Nanoelectronics*. Cambridge University Press (2017).
2. Gao, C., and Xiang, X. -D., "Quantitative Microwave Near-Field Microscopy of Dielectric Properties," *Review of Scientific Instruments 69, pp. 3846-3851* (1998).
3. Zienkiewicz, O. C., *The Finite Element Method in Engineering Science, 3rd ed., pp. 190-191.* McGraw-Hill (1977).
4. Zienkiewicz, O. C., and Taylor, R. L., *The Finite Element Method, 5th ed., Vol. 1: "The Basis," Sections 8.3 and 8.4, pp. 168-172.* Butterworth-Heinemann (2000).
5. Jin, J., *The Finite Element Method in Electro-magnetics, 2nd ed.* Wiley (2002).
6. Fong, J. T., Heckert, N. A., Filliben, J. J., Marcal, P. V., and Rainsberger, R., "Uncertainty of FEM Solutions Using a Nonlinear Least Squares Fit Method and a Design of Experiments Approach," *Proc COMSOL Users' Conf, Oct. 7-9, 2015, Boston, MA,www.comsol.com/ed/direct/conf/conference2015 papers/papers/* (2015).
7. Fong, J. T., Filliben, J. J., Heckert, N. A., Marcal, P. V., Rainsberger, R., and Ma, L. "Uncertainty Quantification of Stresses in a Cracked Pipe Elbow Weldment Using a Logistic Function Fit, a Nonlinear Least Squares Algorithm, and a Super-parametric Method," *Procedia Engineering*, *130, pp. 135-149* (2015).
8. COMSOL, *AC/DC Module User's Guide, Version 5.4, www.comsol.com* (2019).
9. Chollet, F., 2017, *Deep Learning with Python*. Manning Publications (2017).
10. Fong, J. T., Marcal, P. V., Rainsberger, R., Heckert, N. A., and Filliben, J. J., "Design of an intelligent PYTHON code for validating crack growth exponent by monitoring a crack of zig-zag shape in a cracked pipe," *Proc. ASME Pressure Vessels and Piping Division Conf., July 14-19, 2019, San Antonio, TX, U.S.A., Paper PVP2019-93502.* New York, NY: Amer. Soc. Mech. Engineers (2019).
11. COMSOL, *Compiler Module User's Guide, Version 5.4, www.comsol.com* (2019).
12. Filliben, J. J., and Heckert, N. A., 2002, Dataplot: A Statistical Data Analysis Software System, National Institute of Standards & Technology, Gaithersburg, MD 20899, http://www.itl.nist.gov/div898/software/dataplot.html , 2002.

9. Acknowledgment

We wish to thank Drs. **Sergei Yushanov**, **Kyle Koppenhoefer**, **Joshua Thomas**, and **Jeffrey Crompton**, all of Altasim Technologies, LLC, Columbus, OH, for their technical assistance during this investigation. In addition, we wish to thank Altasim Technologies, LLC, for providing the first author of this paper (J. T. Fong) access to a licensed copy of the COMSOL Compiler module for obtaining new research result regarding the conversion of a COMSOL source code (extension .mph) to a compiled application code (extension .exe).

An Assistive Learning Workflow on Annotating Images for Object Detection

Vivian Wen Hui Wong
Engineering Informatics Group
Civil and Environmental Engineering
Stanford University
Stanford, United States
vwwong3@stanford.edu

Max Ferguson
Engineering Informatics Group
Civil and Environmental Engineering
Stanford University
Stanford, United States
maxferg@stanford.edu

Kincho H. Law
Engineering Informatics Group
Civil and Environmental Engineering
Stanford University
Stanford, United States
law@stanford.edu

Yung-Tsun Tina Lee
Systems Integration Division
National Institute of Standards and Technology (NIST)
Gaithersburg, United States
yung-tsun.lee@nist.gov

***Abstract*—We present an end-to-end workflow to generate annotated image datasets for object detection semi-automatically, thereby reducing manual annotation need. With this workflow, which we call assistive learning, we are able to reduce manual annotation time on two experimental datasets by approximately 80%. The experimental results of this work show three contributions of the assistive learning workflow: (1) Savings on human annotation time; (2) generalizability to variable dataset sizes, domains and convolutional neural network (CNN) models; and (3) faster CNN training with limited amount of labeled data using a novel contextual sampling method, thereby a reduction in human workload early on in the assistive learning process. In addition, we wrap the workflow in an interactive annotation interface, allowing annotators without any machine learning experience to speed up the annotation process for training the CNN models.**

Keywords—data annotation workflow, computer vision, object detection, smart manufacturing, vehicle detection

I. Introduction

Object detection is one of the fundamental tasks in computer vision. This task is often useful for industrial applications in a variety of domains, such as medical imaging [1], agriculture [2] and robotics [3]. State-of-the-art object detectors are built with convolutional neural networks (CNNs). With well-built CNN architectures, object detectors can achieve excellent accuracies [4]. However, to obtain a high performance, object detectors must be trained with a vast number of training images. The images are typically labelled with bounding boxes; the process is laborious and time consuming to draw [5]. Given a diverse set of tasks that we would like to use CNNs for, being able to quickly build domain-specific datasets and models is important. There is therefore a need for a faster and more automatic annotation process.

Many existing works have been reported on methods that can reduce data-labeling efforts, including transfer learning [6], semi-supervised learning [7], and weakly supervised learning [8]. However, most of these methods still require a certain amount of training data specific to the problem being solved. Data scientists usually still have to build their own dataset to fine-tune their models. Obtaining ground truth labels for these custom datasets remain crucial and time-consuming.

Another notable technique to reduce training time for building a CNN model is active learning, where the model selects the most informative data points based on certain criteria, then inquires a human annotator to obtain labels of those points [9]. However, existing active learning algorithms face limitations. For instance, most criteria that determine informativeness can only be computed with classification problems and require extra CNN training and inference time to obtain. These limitations are further discussed in Section II.A. With the biggest drawback of active learning being the cost of computation, there is currently no practical, end-to-end workflow that uses active learning algorithms to sample image datasets for object detection.

On the other hand, works have been done to accelerate the obtainment of annotated data. Crowd-sourcing has been proposed to combine the efforts of many annotators [5]. This is, however, still costly when building large, problem-specific datasets that require expert labelers. Self-training, a technique for semi-supervised learning, has been suggested, where a model trained with a small amount of labeled data is employed to generate annotations for the remaining unlabeled data [10]. Most existing self-training works, as discussed later in Section II.B, suffer from the tradeoff between the workload of manual labeling and the accuracy of the predicted labels. The tradeoff is that it would cost more labor in the first place to get a more accurate model, but if we reduce the number of expensive manual annotations, the model trained can be inaccurate, and the generated annotations are of low quality.

In this work, we have developed an assistive learning workflow, which builds upon techniques of active learning and self-training. In active learning, the model samples the most informative data points and sends them to the annotator. In self-training, the model is first trained with partially labeled data, then used to annotate the remaining data. In the proposed assistive learning workflow, the model not only combines the behavior of learners in both techniques, but can also learn from the labeled instances and therefore gradually improve the

Wong, Vivian; Ferguson, Max; Law, Kincho; Lee, Yung-Tsun. "An Assistive Learning Workflow on Annotating Images for Object Detection."
Paper presented at 2019 IEEE International Conference on Big Data, Los Angeles, CA, US. December 09, 2019 - December 12, 2019.

accuracy of predicted labels. This process is iterated through a feedback loop involving the annotator and the model.

This paper describes an end-to-end workflow called assistive learning that consists of a human annotator and a machine annotator cooperating in a feedback loop to annotate images in an object detection dataset. Our work shows that the assistive learning workflow is able to generalize well across object detection models and domains, and to overcome the limitation of labor-accuracy tradeoff. In addition, users can access the workflow using an intuitive user interface (UI), which allows visualization and generation of annotations. To demonstrate the robustness of assistive learning, we conduct experiments on two datasets with two CNN architectures, namely YOLOv3 [11] and Mask R-CNN [12]. The end-to-end workflow reduces the time of annotation by 79.4% and 83.1% in the two experimental datasets. Furthermore, the system is designed such that assistive learning can be utilized without the annotators having to concern themselves with the technical aspects of machine learning.

The rest of the paper is organized as follows. Section II provides an overview of related works. Section III presents a brief introduction to object detection and the CNN architectures deployed in our experiments. Section IV gives the details of the assistive learning workflow. Section V presents the experimental results. Section VI discusses the implications of the results, and, finally, Section VII briefly concludes our work.

II. Related Works

This study relates closely to the existing works on active learning and self-training for object detection. This section discusses how assistive learning leverages these works and the points of departure.

A. Active Learning

The key idea of active learning is the sampling of the most informative data point and querying it to an annotator. Active learning has been shown to reduce labeling costs on image classification tasks [13,14]. Several strategies have been formulated to evaluate the informativeness of a data point.

The most popular active learning sampling strategy for object detection is to use uncertainty indicators as measures of informativeness. The key to uncertainty-based sampling is to first calculate a degree of uncertainty of classification labels using, for example, entropy measure, and then sample the instance with the highest uncertainty. For example, uncertainty-based sampling strategies have been used for object detection tasks with shallow machine learning models [15,16,17]. Other works have applied uncertainty-based sampling to deep neural network architectures [18,19]. The works mentioned above, though achieve good results in many scenarios, sample instances solely based on the classification labels of one or two of the most uncertain detections. More information can be captured by aggregating the scores from each detection to calculate the total uncertainty of one image [20]. However, the method on deep CNN architectures still suffers from two drawbacks: First, the method does not account for the representation of the entire dataset [21]. Second, it has been shown that deep CNN models can fit even random noise labeling with high confidence, thereby reducing the effectiveness of uncertainty calculations [22]. In addition to uncertainty-based sampling, details of other sampling strategies in active learning, such as Query-By-Committee, can be found in [9].

The methods mentioned above mostly evaluate their trained model on every unlabeled instance, then calculate their measure of informativeness based on the outputted classification scores. This is impractical in the annotation scenario, because having a deep CNN model to evaluate every unlabeled image would be costly in both time and computational resources with many unlabeled images. Furthermore, with the most labor-intensive part of annotation being the accurate placement of bounding boxes, a sampling criterion computed with classification scores would not seem to be a natural and practical approach, and do not generalize to single-class object detection datasets.

Besides the above-mentioned active learning methods that solely use classification scores to determine informativeness, it has been proposed to use an intermediate CNN layer as a descriptor function to represent all images in the dataset [23]. The unlabeled image whose representation has the highest average Euclidean distance to all labeled image representations is sampled. This method, although theoretically feasible for CNN architectures and single-class datasets, is computationally inefficient, since every sampling instance requires a new round of training and evaluation with the CNN model. Our sampling method considers not only the average Euclidean distances of unlabeled images, but also some contextual criteria. The contextual sampling strategy is more computationally efficient as it does not require running machine learning models against every image. Furthermore, our method does not rely on the prediction of classification scores to sample the unlabeled images.

B. Self-Training

Many works have used a trained model to complete parts of the annotation process. A model trained with a limited number of labeled images is used to predict labels for a set of either unlabeled or weakly labeled images [24,25].

Some works, in addition to self-training, have involved humans in the process of correcting predicted labels. A Polygon-RNN tool has been introduced to allow the users to correct the polygon segmentation masks being generated by a deep CNN model [26,27]. Hand correction of masks can also be incorporated, and the model is iteratively retrained with corrected labels for medical image analysis [28].

The drawbacks of self-training works mentioned above are that they either do not cover the full end-to-end workflow or focus only on specific datasets. Our work leverages the level of human-machine interaction in existing self-training works by feeding the corrected data back to the model to continuously improve its accuracy throughout the annotation process.

III. Object Detection

Object detection is the task of generating a bounding box around the object of interest [29]. Different object detectors have been reported using machine learning methods [30,31,32]. Nowadays, CNNs (e.g. AlexNet [33], SqueezeNet [34]) trained using deep learning technologies are the predominant approach. There are two main types of CNN architectures for object detectors: one-stage and two-stage.

In this work, we have experimented both a one-stage and a two-stage detector, each with a different dataset, in order to demonstrate that the annotation tool is able to generalize across domains. We implement the one-stage YOLOv3 [11] architecture on a vehicle detection dataset [35], and the two-stage Mask R-CNN architecture on a manufacturing casting defect dataset named GDXray [36]. These two cases were chosen due to the high model scoring accuracies achieved in [35] and [37], as well as the fact that the two datasets have very different types of images. Vehicles on an image are much larger in size than defects on metal castings. Therefore a one-stage detector's grid-based prediction approach can achieve a high accuracy for the vehicle detection problem while dispensing less computational times compared to two-stage detectors. Manufacturing defects, however, are much smaller and can be easily mistaken as noise. Therefore a region proposal stage will increase the detection accuracy. Figs. 1 and 2 illustrate the images of each dataset.

A. *One-stage Object Detectors*

CNN architectures, such as RetinaNet [38], SSD [39] and YOLO [40], are one-stage object detectors whose single task is to directly predict confidence scores for classifying and constructing bounding boxes that surround the objects of interest. One-stage detectors have been shown to train faster than two-stage detectors [11,39].

A popular one-stage object detector is YOLOv3, which is one of the two refined models of YOLO, the first one-stage deep learning object detector published in 2015 [11,40,41]. YOLOv3 has been demonstrated to give a good balance of speed and accuracy in [35] and is adopted as an illustrative implementation of the assistive learning workflow for the vehicle detection dataset.

B. *Two-stage Object Detectors*

Other CNN architectures such as R-CNN [42], Fast R-CNN [43] and Mask R-CNN [12] are two-stage detectors, which consist of an intermediate stage that generates region proposals. Two-stage detectors are able to perform with very high accuracies even for images with fine details but require more computing resources to train [11].

Although training speed differs for different CNN architectures, in general, training good object detectors require a massive amount of training images with ground truth labels.

IV. Assistive Learning

We propose an end-to-end assistive learning workflow. This section is organized into two subsections. In the first subsection, we introduce the underlying framework of assistive learning, and we describe the role of the human annotator and the machine annotator in detail. In the second subsection, we describe the functionalities of an interactive UI, which allow any annotator to engage in the workflow.

A. *End-to-end Workflow*

We introduce a framework that involves the collaboration between a human annotator and a machine annotator, which *assists* and *learns* from the human. Given an unlabeled dataset, the framework consists of the feedback loop shown in Fig. 3.

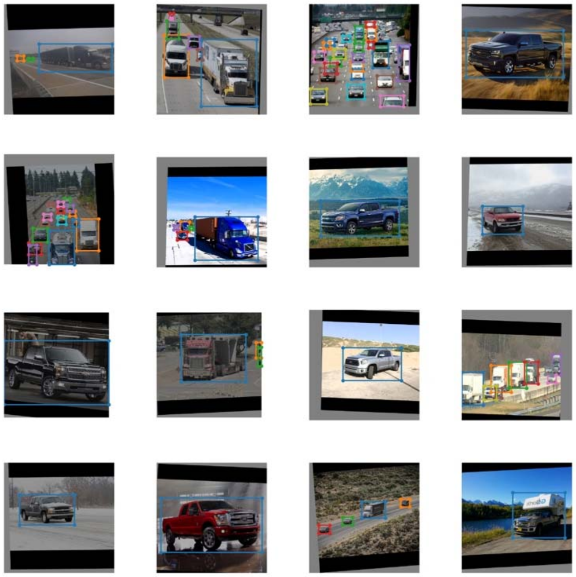


Fig. 1. Examples of labeled vehicle detection images. Vehicles are generally large and easily distinguishable visually.

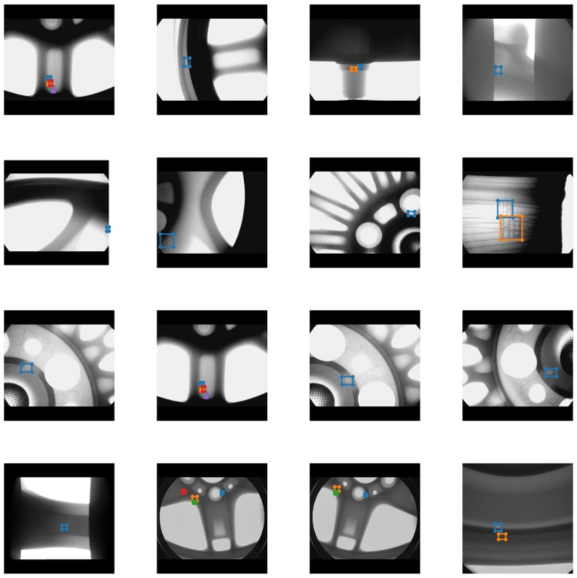


Fig. 2. Examples of labeled GDXray Casting images. Defects are small and very similar to surrounding design features.

The *human annotator* observes and annotates images suggested by the machine annotator. If the machine annotator has already annotated the image, the human annotator reviews the machine-generated labels and correct them if necessary. We consider human-generated labels as ground truth labels. The *machine annotator* consists of two modules: the object detection module and the contextual sampling module. The two modules work together to suggest and label a subsequent image, which is then sent to the human annotator for review.

1. The object detection module is a CNN model that is able to train on the ground truth labels that the human annotator has provided. It can therefore learn from more and more training

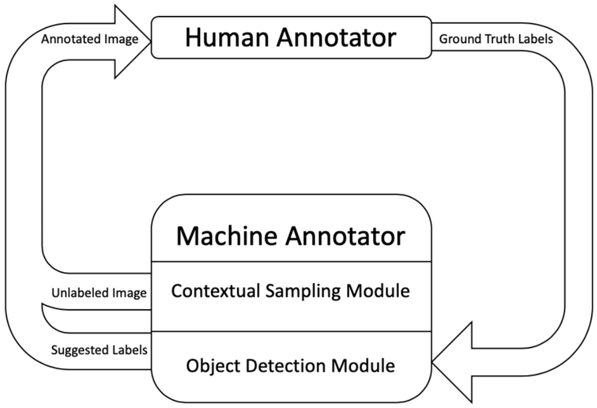


Fig. 3. Assistive learning feedback loop.

samples as we iterate through the loop as shown in Fig. 3. The object detection module annotates the sampled image outputted from the contextual sampling module and suggests the annotation to the human annotator.

2. The contextual sampling module ensures that the next unlabeled image sent to the human annotator satisfies the contextual criteria. The contextual criteria consist of two parts: uniqueness and average Euclidean distance. The criteria are applied on the raw images.

In industry applications, datasets often contain images captured from the same source or specimen. These images are therefore very similar. An example is the GDXray dataset [36], where several images are captured using the same metal casting. In contextual sampling, we avoid repeatedly sampling from the same specimen by examining an image's uniqueness. Uniqueness is defined as follows: Let's define a set of all images in a dataset, $\mathcal{X}$, consisting of labeled images $\mathcal{M}$ and unlabeled images $\mathcal{N}$, i.e. $\mathcal{X} = \mathcal{M} \cup \mathcal{N}$. Suppose there are $K\ (K \geq 1)$ pre-defined specimens in the dataset $\mathcal{X}$, the dataset $\mathcal{X}$ can also be expressed as $\mathcal{X} = \bigcup_{\mathrm{i}=1}^{K} X_i$, where X_i is the set of all images in specimen *i*. An unlabeled image x in specimen *k* (i.e. $x \in \mathcal{N} \cap X_k$, where $1 \leq k \leq K$) is considered unique if specimen *k* has the fewest number of images in the set of labeled images $\mathcal{M}$. In other words, if an unlabeled image x in specimen k ($x \in N \cap X_k$) is considered unique, then the number of images in $X_k \cap \mathcal{M}$ is fewer than or equal to that in all other $X_1, \dots, X_i, \dots, X_K \cap \mathcal{M}$, for $i \neq k$.

The other contextual sampling criterion considers the average Euclidean distance of an image. An image $x \in \mathcal{N}$ is sampled if of all the images that satisfy the uniqueness criterion, *x* has the highest average Euclidean distance to all images in the labeled set $\mathcal{M}$. We denote the average Euclidean distance between an image *x* and all images in $\mathcal{M}$ as $d_x = d(x, \mathcal{M})$.

Following the definition of the two contextual criteria, we now present the procedure of contextual sampling. We calculate the number of images in $X_i \cap \mathcal{M}$ for all $X_i \in \{X_1, \dots, X_K\}$ and select specimen *k* that has the fewest number of images among $X_i \cap \mathcal{M}$. For each unlabeled image $x \in \mathcal{N} \cap X_k$, we compute $d_x = d(x, \mathcal{M})$, the average Euclidean distance between *x* and all images in $\mathcal{M}$ and select the sampled image (denoted by $\hat{x}$) with maximum d_x among all unlabeled images *x* in specimen *k*. The pseudocode of this procedure is presented as shown in Algorithm 1. Note that for images that do not have pre-defined specimens (for example, the vehicle detection dataset), we set the number of specimens $K = 1$. Since there is only one specimen, the average Euclidean distance is the only contextual criterion.

ALGORITHM 1: DETAILED IMPLEMENTATION OF CONTEXTUAL SAMPLING TO SAMPLE THE NEXT IMAGE FROM UNLABELED IMAGES.

```
Input
    $\mathcal{X}$   Set of all images in the dataset
    $\mathcal{M}$   Set of all labeled images in the dataset
    $\mathcal{N}$   Set of all unlabeled images in the dataset
    $K$   Number of specimens in the dataset
    $X_i$   Set of all images from specimen $i$
Set $k \Leftarrow$ specimen with fewest images in $\mathcal{M}$
Set $d_{\hat{x}} \Leftarrow 0$
for $x$ in $\mathcal{N} \cap X_k$ do
    Compute $d_x = d(x, \mathcal{M})$        $\triangleright$ average Euclidean distance between $x$ and all images in $\mathcal{M}$
    if $d_x > d_{\hat{x}}$ then
        Set $d_{\hat{x}} \Leftarrow d_x$ and $\hat{x} \Leftarrow x$
    end if
end for
Output
    $\hat{x}$   Sampled image to be added to $\mathcal{M}$
```

B. *Functionality of User Interface*

A user interface (UI) is necessary for the level of human-machine interaction needed in our workflow. To that end, we develop an interface for the workflow and implement features to visualize, create, edit and remove annotations, to sample images, and to train the object detector.

1) Visualization and Annotation

Fig. 4 illustrates an example of using the interface to visualize bounding boxes generated by the machine annotator. Since the workflow is flexible to the object detector deployed, it can output any box predicted by the object detectors.

Using the library Fabric.js [44], the UI is able to render images and boxes. Furthermore, it is intuitive for a user to rescale bounding boxes with anchor points on the four edges and corners. Deletion of a box is done simply with the click of a key. An additional box can be drawn by pressing down the mouse on the location of one corner and releasing the mouse at the box's diagonal corner. An example of a human-reviewed annotation is illustrated in Fig. 5.

In addition to the intuitive annotating features, a visualization feature of the UI is to allow the annotator to detect with different models. In Fig. 4, we demonstrate that prediction can be done with YOLOv3 and Mask R-CNN on the same image. This is a convenient feature for debugging when switching models, or for the annotator to compare their accuracies.

2) Training

After the human annotator has reviewed and labeled a batch of images, the training function can take the batch of ground truth data and use it to train the object detector. Training will occur whenever the human annotator decides to do so. Since the

Wong, Vivian; Ferguson, Max; Law, Kincho; Lee, Yung-Tsun. "An Assistive Learning Workflow on Annotating Images for Object Detection."
Paper presented at 2019 IEEE International Conference on Big Data, Los Angeles, CA, US. December 09, 2019 - December 12, 2019.

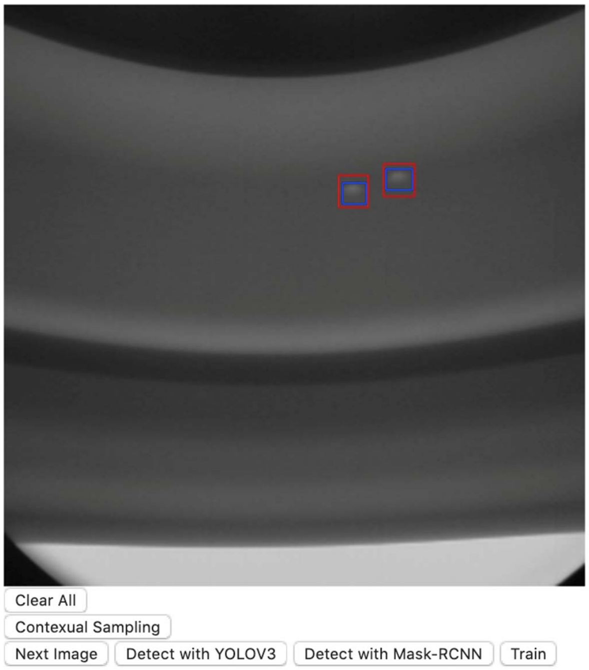


Fig. 4. A screenshot of the annotating UI. Image labeled is from the GDXray Castings dataset. Red bounding boxes are from YOLOv3 object detector. Blue bounding boxes are from Mask R-CNN object detector.

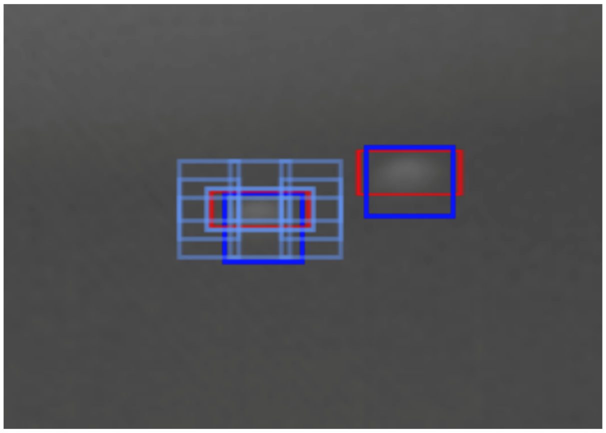


Fig. 5. An example of editing bounding boxes using anchors. With the same image in Fig. 4, the human annotator can manually edit the red bounding boxes to fit them tighter to the defects. The blue bounding boxes recommended by the object detector are kept in place as reference.

model can be trained in the background, the human annotator has the option to do manual annotations while the model trains itself. We specifically allow this parallel process because annotators are usually paid, for example, on an hourly basis, so any unnecessary waiting would be uneconomical.

The training function must be called by the human annotator by pressing the "Train" button. It will not automatically commence whenever an image has been annotated. This is a designed feature, since usually, training after every image is undesirable. Although training time differs for each model, the improved accuracy of adding only one image to the training set does not usually compensate for the cost in training time.

C. *Implementation Details*

The annotation interface developed is an application on the web browser built using the model-view-controller architecture [45]. As shown in Fig. 6, the frontend development of the tool is implemented with Javascript and HTML. The Javascript HTML5 canvas library Fabric.js [44] is used to render the images and bounding boxes on an HTML canvas, as well as to draw, edit or remove bounding boxes. In backend development, the tool uses Python 3.7 for contextual sampling, model training and detection. The CNN implementation is based on the PyTorch framework [46]. A controller module is written to connect the frontend to the backend features.

V. Experimental Results

This section discusses the two case studies to estimate the amount of manual annotation time reduced by assistive learning, and to show that our sampling method is able to reduce the amount of data required for a CNN model to achieve a certain level of accuracy.

A. *Vehicle Detection with YOLOv3*

The YOLOv3 model is trained and evaluated using a total of 3500 vehicle images from [35]. The training set consists of 2600 images, including a combination of daytime camera photos on a highway and online vehicle images. The training set has 4391 vehicle instances (2947 trucks, 344 pick-up trucks, and 1100 cars). The testing set has 900 daytime camera images captured on the highway, and consists of 1226 vehicles (566 trucks, 178 pick-up trucks, and 482 cars). Camera images under various weather and lighting conditions as well as online images are included to ensure that the machine annotator (when both detecting and sampling) can generalize, and that future camera images captured under various conditions can receive an accurate predicted label. We also augment the images by flipping them vertically and horizontally. A weight pre-trained on ImageNet data is used to initialize parameters of the model [29].

1) Annotation Time

To estimate the time taken to annotate, we use the results presented in [5], which states that the average time it takes a person to draw a bounding box is 88.0 seconds [5]. We make the assumption that editing or removing an incorrect bounding box takes half the time, or 44.0 seconds. This is in fact, quite a conservative estimation, since deleting a bounding box takes only a couple of seconds, and the adjustment of an existing box is easy when there is already a suggestion. Using the common standards in the object detection field, an intersection-over-

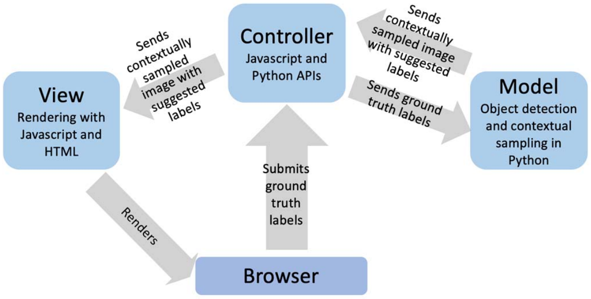

Controller
Javascript and Python APIs
View
Rendering with Javascript and HTML
Model
Object detection and contextual sampling in Python
Browser
Sends contextually sampled image with suggested labels
Sends contextually sampled image with suggested labels
Sends ground truth labels
Submits ground truth labels
Renders

Fig. 6. Model-view-controller design of annotation interface.

union (IOU) of 0.5 or above is considered a correct prediction [47]. We therefore assume that a bounding box will need to be either edited or removed when IOU is under 0.5. Note that we do not count the training time into the calculation of manual annotation time. This is because during training, the human annotator is not required to do any work. Using our system, the human annotator is free to either take a break or annotate while the model trains in the background.

To simulate annotating the vehicle detection dataset with YOLOv3, we split the 2600 images in the training set into 10 smaller datasets as shown in Table I, with dataset sizes ranging from the smallest to the largest. The order of images is obtained using the contextual sampling method described earlier in Section IV. Since the vehicle detection dataset does not contain specimens, we use average Euclidean distance as the only contextual criterion. The first dataset is considered fully annotated by a human annotator. For the first dataset, we train it for 5 epochs with weights frozen for intermediate layers of the CNN, and another 35 epochs without holding weights fixed. We test the newly trained model on the second dataset, and compare the evaluation result to ground truth labels to compute the amount of human annotation time needed to correct inaccurate labels. The second dataset, after the simulated correction process, is now considered fully annotated. The model then trains on all fully annotated images, which now consist of the first and the second dataset, for 35 epochs. The process is repeated for all remaining datasets: For each dataset after the first, we evaluate it with the model trained with previously added fully annotated images. We then add the evaluated images to the fully annotated images, simulating the hand-correction of a human annotator, then train for 35 epochs.

Since the goal of this study is to demonstrate the effectiveness of the techniques proposed, no hyperparameter optimization is conducted. The training schedule is also not aiming to fully minimize training loss. As shown in Table I, when trained on the same training set and tested on the testing set as [35], the model achieves an mean average precision (mAP) of 0.831, which is 88.7% of the mAP of 0.937 in [35]. Our accuracy is sufficiently high for the purpose of this study, but to achieve accuracies as high as [35], one will need to tune all hyperparameters and train for more epochs.

TABLE I. TOTAL ANNOTATION TIME TO LABEL ALL IMAGES USING YOLOV3 WITH VEHICLE DETECTION DATASET. RESULTS COMPARED WITH AND WITHOUT THE USE OF MACHINE ANNOTATOR.

# of images trained	**# of images annotated with machine annotator**	**Time with human annotator alone (hours)**	**Time with assistive learning (hours)**	**% Time reduced**	**mAP**
30	5	0.54	0.48	11.4%	0.042
35	10	0.37	0.32	13.3%	0.197
45	25	1.10	0.88	20.0%	0.097
70	30	1.54	0.83	46.0%	0.343
100	100	6.43	2.86	55.5%	0.227
200	200	9.34	2.59	72.3%	0.535
400	400	15.11	3.37	77.7%	0.674
800	800	29.02	5.15	82.3%	0.691
1600	1000	37.74	6.28	83.4%	0.767
2600	900	32.24	4.68	85.5%	0.831
Total hours	-	133.42	27.44	**79.4%**	

Table I shows that when annotating the entire vehicle detection dataset, assistive learning significantly reduces manual annotation time compared to a human annotator working alone. As more labeled images are added to the training set, the machine annotator saves a higher percentage of time for the human annotator, due to improved performance of the machine annotator with more training.

2) Sampling Methods

By experimenting with YOLOv3 on the vehicle detection dataset, we compare the contextual sampling method (average Euclidean distance as the only contextual criterion) proposed in Section IV with random sampling. To conduct the experiment, we train a YOLOv3 model on various numbers of images, then compare the results with those obtained from random sampling. Pre-trained ImageNet weights are used for training. Each time the model is trained on 5 epochs with weights of intermediate layers fixed, then 35 epochs not fixed. Five identical, independent experiments are run.

Fig. 7 shows that the contextual sampling method improves the performance of the object detector when there are little labeled data. With 50 labeled images, the mean mAP from 5 independent runs is 0.32 for random sampling and 0.40 for contextual sampling. Contextual sampling improves the mean mAP by 25.0%. With 100 labeled images, contextual sampling improves the mean mAP by 23.3%, from 0.43 with random sampling to 0.53 with contextual sampling. This is because with contextual sampling, the labeled data show less similarity. The object detector can therefore generalize and fit faster. As the number of labeled images grows larger, however, the performance of the model given the two sampling methods converge.

B. *Castings Defect Detection with Mask R-CNN*

The Mask R-CNN model is trained and evaluated using images from the GDXray Castings dataset [36]. The Casting Series dataset contains 2727 images. Many of the images in the

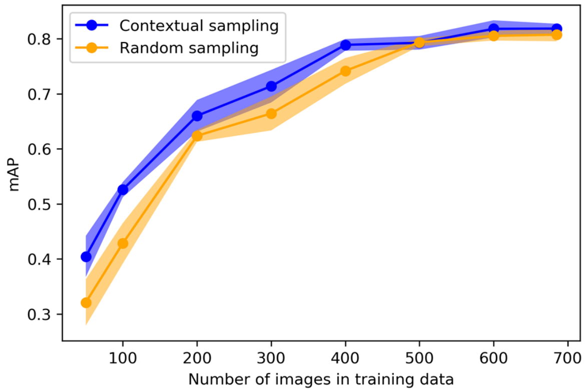


Fig. 7. Relationship between mean average precision (mAP) and number of training images, sampled randomly and contextually on the vehicle detection dataset. mAP are obtained by testing on the same test dataset. The shaded areas show the standard deviations from 5 independent runs.

TABLE II. TOTAL ANNOTATION TIME TO LABEL ALL IMAGES USING MASK R-CNN WITH GDXRAY CASTINGS DATASET. RESULTS COMPARED WITH AND WITHOUT THE USE OF MACHINE ANNOTATOR.

# of images trained	**# of images annotate d with machine annotator**	**Time with human annotator alone (hours)**	**Time with assistive learning (hours)**	**% Time reduced**	**mAP**
30	5	0.22	0.09	61.1%	0.483
35	10	0.46	0.28	39.5%	0.150
45	25	2.47	0.31	87.6%	0.616
70	30	2.98	0.55	81.6%	0.750
100	100	8.75	1.14	87.0%	0.656
200	200	16.57	3.24	80.5%	0.793
400	285	23.44	4.11	82.5%	0.768
685	171	12.27	1.64	86.7%	0.872
Total hours	-	67.17	11.34	**83.1%**	

dataset were unlabeled or had ambiguous labels, so we chose to train on the 685 training images specified in [37,48]. Testing was conducted on the 171 images specified in [37,48]. This dataset contains a single casting defects class, but the labeled X-ray images come from 33 specimens. Images coming from the same specimen are very similar. The model parameters are initialized with weights pre-trained on COCO data [49].

1) Annotation Time

Similar to the vehicle detection problem, to simulate annotating the GDXray Castings dataset with Mask R-CNN, we split the 685 images in the training set into 8 smaller datasets, with dataset sizes ranging from the smallest to the largest. Contextual sampling is applied (with average Euclidean distance and specimen uniqueness as the contextual criteria). Due to high computing time of Mask R-CNN, we only train 5 epochs for each dataset. The first dataset includes an additional epoch for fixing the weights on all except the output layer. We evaluate every dataset after the first dataset with the model trained with previously evaluated datasets.

Table II shows the reduction of manual annotation time using a Mask R-CNN object detector to assist the annotation of GDXray Castings images. In this experiment, the models are trained on 1 epoch with fixed weights on intermediate layers followed by 5 epochs without holding weights fixed, the mAP of 0.872 for constructing the bounding boxes surrounding the defects is obtained using the same 685 training images and 171 testing images as discussed in [37,48]. The result is also compatible with the CNN model training under similar setting [37]. Although it is not conducted in this study because of extensive computational time and hyperparameter tuning required, as reported in [37], an mAP of 0.957 could be obtained when trained with 80 epochs of fixed weights on intermediate layers and another 80 epochs without holding weights fixed. That is, the trained model in this experiment achieves 91.1% of the best mAP as reported in [37]. This result is reasonable, considering that the model parameters are not optimally updated at each incremental training step. In short, the results show the workflow has the capability to greatly reduce manual annotation time.

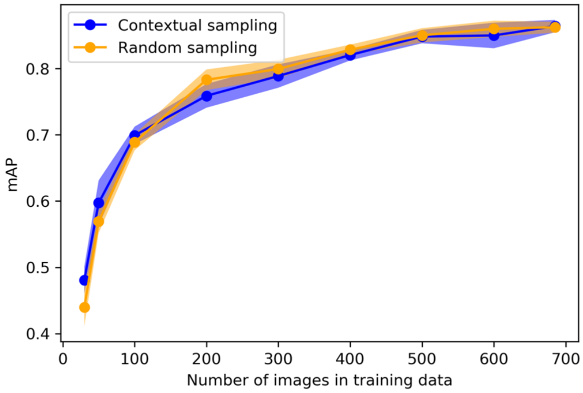


Fig 8. Relationship between mean average precision (mAP) and number of training images, sampled randomly and contextually on the GDXrays castings dataset. mAP are obtained by testing on the same test dataset. The shaded areas show the standard deviations from 5 independent runs.

2) Sampling Methods

Since the GDXray Castings dataset consists of specimens that consist of similar images, the contextual sampling method considers uniqueness and average Euclidean distance when sampling. Similar to the sampling experiments conducted in Section A, with GDXray dataset and Mask R-CNN object detector, similar results are obtained as shown in Fig. 8. When the dataset size is small, contextual sampling slightly outperforms random sampling by 9.4% mAP when trained on 30 images and 5.0% mAP when trained on 50 images.

Even though the contextual sampling method improves performance with limited labeled images, the human annotator should still take into consideration the labor-accuracy tradeoff. More upfront labor investment leads to better predicted labels. However, with the assistive learning workflow, the tradeoff can be converted to a training time-labor tradeoff. The human annotator can choose to train more frequently to increase accuracy of predicted labels, thereby reducing data-labeling efforts.

VI. DISCUSSION

In this section, we evaluate our workflow based on the experimental results presented in the previous section. Besides having shown that the workflow efficiently reduces the workload of human annotators, this section discusses additional benefits and the costs to consider when using assistive learning.

A. Generalization on Datasets

By experimenting our tool with two datasets, we were able to show that the workflow can be applied to problems from different domains due to the flexibility of the object detector and sampling strategy. In this subsection, we outline why the two datasets differ a lot in object detection, in order to illustrate the importance of cross-domain capability.

The two datasets studied are both commonly observed industrial problems in engineering yet differ mainly in the following ways:

- **Size of objects** Vehicles usually take up a larger portion of an image compared to casting defects. YOLO is not ideal for small objects since it limits the number of nearby detections [50]. Two-stage architectures that use regional proposal are shown to be better at locating small objects [40]. Alternatively, one-stage architectures can be finetuned to target small objects [51].

- **Noisiness of background** Vehicles, which have distinct colors and shape, are normally quite distinguishable from their backgrounds. On the other hand, casting defects often look very similar to design features like holes and edges [37]. CNN architectures with high quality feature maps would be better at distinguishing such defects.
- **Classes and specimens** The vehicle detection dataset has three classes of vehicles, whereas the GDXray Castings dataset has a single class but contains similar images from the same specimens. It is therefore important to have a contextual sampling method that is applicable to either scenarios.

In summary, it is important for the human annotator to take caution when choosing a suitable object detector (i.e. the CNN architecture) for annotation, as it greatly depends on the type of objects in the custom image dataset being built. Our workflow offers flexibility to customize CNN models and datasets, and introduces a computationally efficient contextual sampling method that is applicable to a variety of datasets.

B. Costs for Consideration

Although the tool is able to reduce the need of manual labor, there are additional costs to be considered. Firstly, the cost of training time can be quite significant with more complex CNN models. With these complex models, it might cost less time overall to annotate with human annotators. For example, with a one-stage object detector like YOLOv3, training time increases linearly with the number of training images. Although training the object detector more often will allow it to learn faster and reduce labeling efforts, one must evaluate whether the overall cost in time is more expensive than hand-annotation.

The second factor is the monetary cost of computational resources. CNN model training nowadays is mostly done on GPUs, which are powerful in computation. For the tool to train effectively, it therefore must connect to either a cloud GPU or a physical GPU, which can be costly. To find the most economical way to annotate a dataset, it may worth to compare the cost of hiring human annotators versus the available (and, possibly, purchasing) computing resources.

C. Future Work

The assistive learning workflow presented in this work is able to reduce manual annotation time and can be applied to annotate and build datasets. However, the training process of the machine annotator can be computationally expensive. Future work could explore training time and its tradeoff between prediction accuracy. It would also be interesting to further evaluate the workflow on object detection with other datasets and to apply the approach to other applications, such as image segmentation, natural language processing and multimodal deep learning.

VII. Conclusion

This paper has presented an assistive learning workflow to annotate bounding boxes on images for the task of object detection. By conducting two experiments, we are able to show that by connecting a human annotator and a machine annotator with a feedback loop, we can significantly reduce the amount of manual annotation time. We also propose a novel contextual sampling method, which improves the performance of YOLOv3 and Mask R-CNN object detector when trained on a very limited number of labeled images. Together, we deploy the workflow in the backend of an intuitive user interface. The experimental results show that our workflow is able to generalize to datasets in different domains and is flexible to various object detection architectures. We also note that there exists a tradeoff between training time and hand-annotation workload.

In summary, this work proposes an end-to-end workflow that uses assistive learning to annotate images for object detection. Our estimation of savings in hand-annotation efforts is conservative yet still exceptional. We hope that this workflow can be adopted to produce valuable labeled datasets across all domains, while minimizing manual labor.

VIII. Acknowledgment and Disclaimer

This research is partially supported by the Smart Manufacturing Systems Design and Analysis Program at the National Institute of Standards and Technology (NIST), US Department of Commerce, Grant Numbers 70NANB18H193 and 70NANB19H097 awarded to Stanford University.

Certain commercial systems are identified in this article. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

IX. References

[1] Z. Li, M. Dong, S. Wen, X. Hu, P. Zhou, and Z. Zeng, “CLU-CNNs: Object detection for medical images,” *Neurocomputing*, vol. 350, pp. 53–59, Jul. 2019.

[2] L. Saxena and L. Armstrong, “A survey of image processing techniques for agriculture,” *Proceedings of Asian Federation for Information Technology in Agriculture*, pp. 401-413, 2014.

[3] D. Xu, Q. Huang and H. Liu, "Object detection on robot operation system," *2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1155-1159, 2016,

[4] Z. Zou, Z. Shi, Y. Guo, and J. Ye, “Object detection in 20 years: a survey,” *arXiv preprint arXiv: 1905.05055*, 2019.

[5] H. Su, J. Deng, and F-F. Li, “Crowdsourcing annotations for visual object detection,” *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence,* 2012.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, 2010.

[7] X. Zhu, "Semi-supervised learning literature survey," Technical Report TR-1530, Univ. of Wisconsin-Madson, 2005.

[8] Z-H. Zhou, “A brief introduction to weakly supervised learning,” *National Science Review*, vol. 5, no. 1, pp. 44–53, 2018.

[9] B. Settles, "Active learning literature survey", University of Wisconsin-Madison Department of Computer Sciences, Technical Report 1648, 2009.

[10] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods", *Proc. 33rd. Annual Meeting of the Association of Computational Linguistics*, pp. 189-196, 1995.

[11] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[12] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp. 2980-2988, 2017.

[13] A. Kapoor, K. Grauman, R. Urtasun and T. Darrell, "Active learning with gaussian processes for object categorization," *2007 IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, Brazil, pp. 1-8, 2007.

[14] A. J. Joshi, F. Porikli and N. Papanikolopoulos, "Multi-class active learning for image classification," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, pp. 2372-2379, 2009.

[15] Bietti, A. “Active learning for object detection on satellite images,” Technical Report, Caltech, 2012. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.1233&rep=rep1&type=pdf (Accessed: August 1, 2019).

[16] Y. Abramson, Y. Freund, “Active learning for visual object detection,” San Diego:Department of Computer Science and Engineering, University of California, 2006.

[17] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," *IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, pp. 1449-1456, 2011.

[18] D. Wang and Y. Shang, "A new active labeling method for deep learning," *2014 International Joint Conference on Neural Networks (IJCNN)*, Beijing, China, pp. 112-119, 2014.

[19] C. Feng, M.-Y. Liu, C.-C. Kao, T.-Y. Lee, "Deep active learning for civil infrastructure defect detection and classification", *ASCE International Workshop on Computing in Civil Engineering*, Seattle, WA, USA, pp. 298-306, 2017.

[20] C-A. Brust, C. Käding, and J. Denzler, “Active learning for deep object detection,” *arXiv preprint arXiv:1809.09875*, 2019.

[21] T. He *et al.*, “An active learning approach with uncertainty, representativeness, and diversity,” *The Scientific World Journal.*, vol. 2014, Article ID 827586, 6 pages, 2014.

[22] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” *International Conference on Learning Representations,* vol. abs/1611.03530, 2017

[23] A. Smailagic *et al.*, “MedAL: Deep active learning sampling method for medical image analysis,” *arXiv preprint arXiv:1809.09287*, Sep. 2018.

[24] B. Adhikari, J. Peltomaki, J. Puura and H. Huttunen, "Faster bounding box annotation for object detection in indoor scenes," *2018 7th European Workshop on Visual Information Processing (EUVIP)*, Tampere, Finland, pp. 1-6, 2018.

[25] C. Rosenberg, M. Hebert and H. Schneiderman, "Semi-supervised self-training of object detection models," *2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1*, Breckenridge, CO, USA, pp. 29-36, 2005.

[26] L. Castrejón, K. Kundu, R. Urtasun and S. Fidler, "Annotating object instances with a Polygon-RNN," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 4485-4493, 2017.

[27] D. Acuna, H. Ling, A. Kar and S. Fidler, "Efficient interactive annotation of segmentation datasets with Polygon-RNN++," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 859-868, 2018.

[28] B. Lutnick *et al.*, “Iterative annotation to ease neural network training: Specialized machine learning in medical image analysis.,” *Nature Machine Intelligence*, vol. 1, no. 2, pp. 112-119, 2019.

[29] O. Russakovsky *et al.*, “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, pp. 1-42, 2015.

[30] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Kauai, HI, USA, pp. I-I, 2001.

[31] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, vol. 1, pp. 886-893, 2005.

[32] P. Felzenszwalb, D. McAllester and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," *2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, pp. 1-8, 2008.

[33] M. Z. Alom *et al.*, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.

[34] F. N. Iandola, *et al.*, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," *arXiv preprint arXiv:1602.07360,* 2016.

[35] R. Hou, S. Jeong, K. H. Law, and J. P. Lynch, “Reidentification of trucks in highway corridors using convolutional neural networks to link truck weights to bridge responses,” *Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring*, St. Louis, Missouri, USA, 2019.

[36] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, H. Lobel, and M. Carrasco, “GDXray: The database of X-ray images for nondestructive testing,” *Journal of Nondestructive Evaluation*, vol. 34, no. 4, p. 42, Nov. 2015.

[37] M. Ferguson, R. Ak, Y.-T. T. Lee, and K. H. Law, “Detection and segmentation of manufacturing defects with convolutional neural networks and transfer learning,” *Smart and Sustainable Manufacturing Systems*, vol. 2, no. 1, pp. 137-164, 2018.

[38] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2999-3007, 2017.

[39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single shot multibox detector,” *European conference on computer vision*, pp. 21--37, 2016.

[40] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, real-time object detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 779-788, 2016.

[41] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 6517-6525, 2017.

[42] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 580-587, 2014.

[43] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, 2015, pp. 1440-1448.

[44] Printio.ru Lab, “Fabric.js.” Available: http://fabricjs.com. [Accessed: 1-Jul-2019].

[45] J. Deacon, “Model-view-controller (MVC) architecture,” *Computer Systems Development*, pp. 1-6, 2005.

[46] A. Paszke *et al.*, “Automatic differentiation in PyTorch,” in *31st Conference on Neural Information Processing Systems (NIPS)*, pp. 1-4, 2017, [online] Available: pytorch.org.

[47] A. Arnab and P. H. S. Torr, "Pixelwise instance segmentation with a dynamically instantiated network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 879-888, 2017.

[48] M. Ferguson, R. Ak, Y. T. Lee and K. H. Law, "Automatic localization of casting defects with convolutional neural networks," *2017 IEEE International Conference on Big Data (Big Data)*, Boston, MA, pp. 1726-1735, 2017.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Dolla ́r, and C. L. Zitnick. Microsoft Coco: Common Objects in Context. *arXiv preprint arXiv:1405.0312*, 2014.

[50] Q. Peng *et al.*, "Pedestrian detection for transformer substation based on gaussian mixture model and YOLO," *8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, Hangzhou, China, pp. 562-565, 2016.

[51] G. Cao *et al.*, “Feature-fused SSD: fast detection for small objects,” in *9th International Conference on Graphic and Image Processing*, Qingdao, China, volume 10615, page 106151E, International Society for Optics and Photonics, 2018.

Chemical Structure of Medium-Scale Liquid Pool Fires

Ryan Falkenstein-Smith*, Kunhyuk Sung, Jian Chen, and Anthony Hamins*

National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, Maryland, USA
ryan.falkenstein-smith@nist.gov and anthony.hamins@nist.gov

*Corresponding authors

Highlights:

- Gas species and soot measurements of steadily burning methanol, ethanol, and acetone 30 cm pool fires are reported.
- Measurements were verified using theoretical gas species ratios.
- Flame structure was found to be in agreement with previous work.
- Gas species concentrations were shown to vary between different fuels.

Abstract: This work documents a series of time-averaged local gas species measurements made throughout the centerline profile of 30 cm methanol, ethanol, and acetone pool fires steadily burning in a quiescent environment. All gas species measurements were obtained using extractive sampling, then analyzed using a gas chromatograph equipped with a mass spectrometer detector (GC/MS). The volume fraction of each species was calculated via the number of moles identified by the GC/MS at each location along the fire's centerline. Soot mass fractions were also measured during the gas-sampling process. The gas species and soot mass fractions were compared at different locations for fires burning three fuels.

Keywords:

Moderately-sized pool fires, acetone fuel, ethanol fuel, methanol fuel, gas species measurements

1. Introduction

Computational fire models, such as the Fire Dynamic Simulator (FDS) [1], have increased in usage within the fire protection engineering community due to their effectiveness, ease of use, and the decreased cost of computational power. To be reliable, fire models require validation to aid in their development and confirm their accuracy and limitations. The objective of this research is to provide experimental data for use in fire model development and evaluation.

Pool fires are a fundamental combustion configuration of interest. In a pool fire, the fuel surface is flat and horizontal, which provides a simple and well-defined configuration for testing models and furthering the understanding of fire phenomena. In moderate and large-scale pool fires, radiative heat transfer is the dominant mechanism of heat feedback to the fuel surface. Species concentrations and temperatures control the radiative heat transfer. An area of particular interest is the fuel rich core just above the pool surface, where vaporizing fuel molecules and other gas

Falkenstein-Smith, Ryan; Sung, Kunhyuk; Chen, Jian; Hamins, Anthony. "Chemical Structure of Medium-Scale Liquid Pool Fires." Paper presented at 13th International Symposium on Fire Safety Science, Waterloo, CA. April 27, 2020 - May 01, 2020.

species can absorb energy that otherwise would have been transferred to the fuel surface. An essential component in this zone is the spatial distribution of gas-phase chemical species. It is well known that the local gas species within a fire is a critical component of its structure. However, there are few studies in the pool fire literature that have reported local chemical species measurements, which elucidate the chemical structure of the fire and provide insight on its kinetic, heat, and mass transfer processes.

The purpose of this study is to characterize the spatial distribution of stable gas-phase chemical species in moderate-scale liquid pool fires steadily burning in a well-ventilated, quiescent, environment. Here, methanol, ethanol, and acetone are the fuels of interest. Contrary to ethanol and acetone, fires established using methanol are unusual as no carbonaceous soot is present or emitted.

In this study, all gas species measurements are made in a 30 cm diameter pool fire using liquid fuels. These fuels are selected since the measurements complement results from previous studies, including analyses of the mass burning rate, the temperature and velocity fields, radiative emission, flame height, and pulsation frequency [2, 3, 4, 5]. Additional characterization of these fires enables a more comprehensive understanding of its detailed structure, enhancing the understanding of fire physics.

2. Experimental Methods

2.1 Pool Burner Setup

A circular, stainless-steel pan with an outer diameter of 30 cm, 15 cm deep, and a 0.16 cm wall thickness was used as the pool burner. As shown in Fig. 1, the burner has an overflow basin, which extended 3.0 cm beyond the burner wall. The burner is fitted with legs such that the burner rim was 30 cm above the ground. The bottom of the burner was maintained at a constant temperature by flowing water (20 °C ± 3 °C) through the 3 cm section on the bottom of the fuel pan. Additionally, a fuel level indicator was positioned near the center of the burner to monitor the fuel level during burning.

The pool burner was located under a canopy hood surrounded by a 2.5 m x 2.5 m x 2.5 m enclosure made of a double-mesh screen wall. The walls of the enclosure were formed by a double layer wire-mesh screen (5 mesh/cm) to reduce the influence of ambient air flows that could disrupt the flow field. All measurements were made once the mass burning rate reached steady-state, achieved approximately 10 min after ignition.

The time-averaged mass burning rate, $\dot{m}$, was determined from the rate at which fuel was delivered to the pool from a reservoir positioned on a mass load cell located outside the enclosure and monitored by a data acquisition system (DAQ). The operator was able to observe a close up of a slightly discernible dimple on the fuel surface using a live video feed. The fuel level was maintained at 10 mm below the burner rim by manually adjusting the fuel flow with a needle valve.

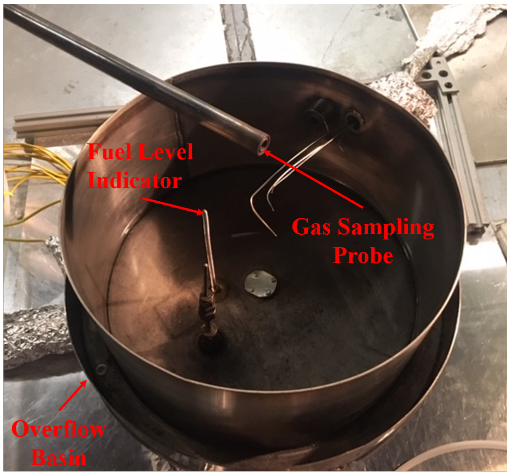


Fig. 1. 30 cm pool burner with fuel level indicator, overflow section, and water-cooled gas sampling probe.

The idealized heat release rate of each fuel, $\dot{Q}$, was calculated from Eq. 1 using the time-averaged mass loss measurements:

$$\dot{Q} = \dot{m}\,\Delta H_c \tag{1}$$

where ΔH_c is the heat of the combustion of the burned fuel provided by DIPPR® [7] and listed in Table 1.

The mean flame height was estimated from 3600 frames obtained from high-quality video recordings of the pool fire experiments, using a technique reported in Ref. [8]. Frames were processed using MATLAB's Image Processing Toolbox[1]. Imported RGB images were decomposed into binary (black and white) images using a pre-set threshold level. The flame height for a single frame was defined as the distance between the pool surface and flame tip established using MATLAB software. All measurements were repeated, then averaged to provide mean values.

2.2 Temperature Measurements

Time-averaged temperature measurements were made along the pool fire centerline. The height locations ranged from 2 to 60 cm relative to the burner rim for all fires. Additional measurements were made higher, up to 100 cm, in the acetone fire to sample its taller flame. A fine S-type thermocouple with a diameter of 50 µm (P10R-001, OMEGA) was positioned onto a traverse

[1] Certain commercial products are identified in this report to specify adequately the equipment used. Such identification does not imply a recommendation by the National Institute of Standards and Technology, nor does it imply that this equipment is the best available for the purpose.

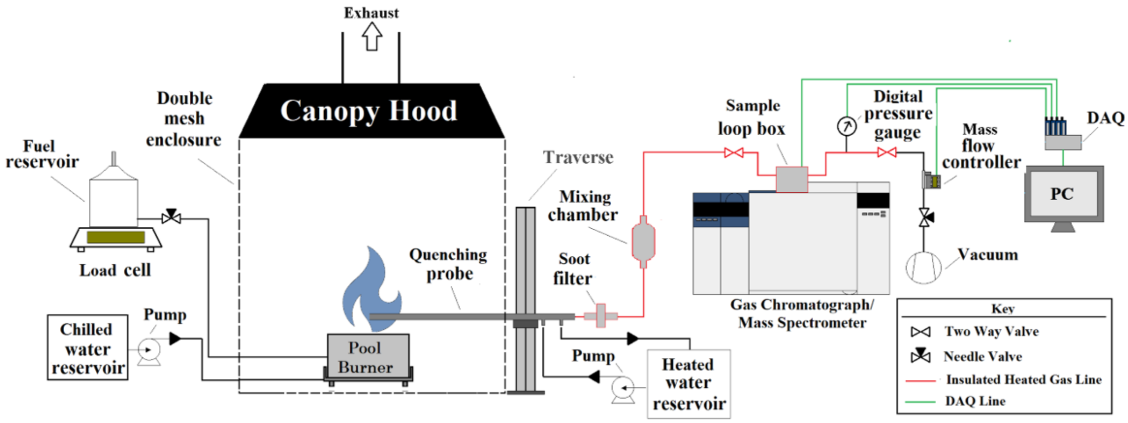


Fig. 2. A schematic of the extractive sampling setup used to extract and transport fire samples from the pool fire to the GC/MS.

such that the bare thermocouple bead was centered above the middle of the burner. The traverse controlled the position of the thermocouple relative to the burner rim as prompted via computer. Temperature measurements were sampled at 250 Hz for 2 min which represented more than 300 fire pulsing cycles [9]. The thermal inertia and radiative heat loss associated with the thermocouple were corrected following Shaddix [10]. The temperature dependent emissivity of platinum was taken from Ref. [11]. The thermal inertia correction had little influence ($<$ 5 K) on the mean temperature but significantly altered the RMS.

2.3 Measuring the Volume Fraction of Gas Species

Figure 2 displays the flow diagram for gas sampling into an Agilent 5977E Series Gas Chromatograph equipped with a mass spectrometer detector (GC/MS) fitted with a thermal conductivity detector (TCD). After achieving steady-state burning conditions, approximately 10 min after ignition, flow was initiated by a vacuum pump located downstream from the GC/MS was initiated. Gas samples were collected using a water-cooled probe. The probe was composed of two concentric, stainless-steel tubes with outer annular coolant flow and inner, extracted, gas-sample flow. The inner and outer tube diameters were 8 mm and 16 mm, respectively. Water at 90 °C flowed through the sampling probe during the experiment. The remainder of the sampling line leading into the GC/MS was heated with electrical heating tape to 140 °C to prevent condensation of water and fuels in the line.

The gas sampling period varied from 12 to 25 min, depending on the sampling location within the fire. Ensuring that the gas sample had completely swept through the GC/MS sample loop, the sampling flow was controlled using a mass-flow controller (Alicat Scientific MC-Series) located in front of the vacuum pump within the sampling line. During the gas sampling procedure, the volumetric flow was approximately 0.2 SLPM, recorded using a DAQ at 2.0 Hz. The mass-flow controller also provided temperature readings of the gas flow.

After the gas sampling period, two quarter-turn valves located on opposite ends of the GC/MS sample loop within the sampling line were closed. Once the sampled gas reached equilibrium, pressure measurements, obtained from a digital pressure gauge (OMEGA DPG409-030DWU), and temperature measurements, acquired by a K Type Thermocouple located at the GC/MS sample loop injection port, were collected at 2.0 Hz for 50.0 s. After collecting pressure and temperature measurements, the sampled gas was injected into the GC/MS.

The volume fraction, $\bar{X}_i$, was calculated from the ratio between the number of moles of a given gas species, n_i, and the total number of moles, n_{tot}. The moles of a given species were identified using the mass spectrometer and quantified from the TCD within the GC/MS. The total number of moles was determined from the summation of moles for each species detected by the TCD.

$$\bar{X}_i = \frac{n_i}{n_{\text{tot}}} \tag{2}$$

The mass fraction, $\bar{Y}_i$, of each species i was calculated from the measured volume fraction, $\bar{X}_i$, using the following expression:

$$\bar{Y}_i = \frac{\bar{X}_i \, W_i}{W_{\text{tot}}} \tag{3}$$

where W_i is the molecular weight of a given species and W_{tot} is the average molecular weight of the sample represented by

$$W_{\text{tot}} = \sum \bar{X}_i \, W_i \tag{4}$$

All measurements using the GC/MS were repeated at least twice along the centerline of the pool fire, at the same positions as the temperature measurements. Gas species concentration measurements made at the same location were averaged. The variance in the gas species volume fraction was a function of position and species.

2.4 Soot Mass Fraction Measurements

Soot was collected simultaneously with gas samples using the sampling procedure described in Section 2.3. Before a test, a desiccated 47.0 mm diameter polytetrafluoroethylene (PTFE) filter was weighed and placed into an in-line stainless steel particulate filter holder. During an experiment, the filter holder was positioned within the gas sampling line behind the sampling probe and heated to 140 °C, using heating tape to prevent condensation of water and liquid fuels on the filter. After testing, the PTFE filter was removed from the filter holder and dried in a desiccator. After desiccating for 48 h, the PTFE filter's final weight was measured. Typically, 2 mg of soot was collected during the sampling period. To obtain a meaningful sample, the sampling period varied from 12 min to 25 min depending on the sampling location within the fire.

After some tests, soot deposits were observed on the inner walls of the quenching probe. Desiccated gun cleaning patches were used to collect soot on the inside of the sampling probe. At least two patches were used to collect soot on the inside of the probe. Soot collection on the inside of the probe concluded once a used patch was observed to have no soot. Patches were weighed immediately before and 48 hrs after cleaning the inside of the probe. The soot collected from the dry patches was accounted for when calculating the soot mass fraction. The portion of the soot collected on the inner walls of the quenching probe relative to the PTFE filter varied based on the sampling location. The mass of the PTFE filter and cleaning patches were measured three times before and after each test.

The soot mass fraction, Y_s, was computed from the mass of the soot collected from the PTFE filter and gun cleaning patches, m_s, the ratio of the mass-flow controller's temperature reading, T_∞, to the temperature at the probe entrance, T_g, the total mass of gas sampled, m_t, based on the mass-flow controller readings:

$$Y_\mathrm{s} = \frac{m_\mathrm{s}}{m_\mathrm{t}}\frac{T_\infty}{T_\mathrm{g}} \tag{5}$$

The total mass of gas sampled was estimated from the product of the average volumetric flow rate measured by the mass-flow controller, $\dot{V}$, the density of the sample gas injected into the GC/MS, ρ_g, and the gas sampling time, δt.

$$m_t = \dot{V}\rho_\mathrm{g}\delta t \tag{6}$$

In Eq.6, the density of the sample gas was determined from the total mass detected in the TCD chromatogram, m_{tot}, for the injected sample volume, V_s.

$$\rho_\mathrm{g} = \frac{m_\mathrm{tot}}{V_\mathrm{s}} \tag{7}$$

2.5 Uncertainty Analysis

The expanded uncertainties of the mass burning rate, mean flame height, volume fraction of gas species, and soot mass fractions were estimated through a combination of Type A and B evaluation of standard uncertainty using a 95 % confidence interval and a coverage factor of 2. The Type A evaluation of standard uncertainty was determined from the variance of repeated measurements. The Type B evaluation of standard uncertainty was defined as the bias errors in the instrumentation. The combined uncertainty of calculated parameters were estimated using the law of propagation of uncertainty. The uncertainty of all measurements is discussed further in Ref. [6].

3. Results

3.1 Flame Observations

Figure 3 displays a series of snapshots depicting the puffing cycle for methanol, ethanol, and acetone pool fires. A repeated cycle was observed in each of the pool fires; uniformly curved flame sheets present at the burner rim repeatedly rolled towards the fire centerline to form a long and narrow plume.

The shape and visible color of the fires differed between fuel types. The methanol fire appeared to be completely blue, whereas the ethanol, and acetone fires were luminous and yellow. The methanol pool fire was observed to exhibit a quasi laminar flame structure compared to the more turbulant nature of fires seen for the other two fuels. The observed dynamic shapes were consistent with previous experiments [3, 4, 5, 12, 13, 14].

Falkenstein-Smith, Ryan; Sung, Kunhyuk; Chen, Jian; Hamins, Anthony. "Chemical Structure of Medium-Scale Liquid Pool Fires." Paper presented at 13th International Symposium on Fire Safety Science, Waterloo, CA. April 27, 2020 - May 01, 2020.

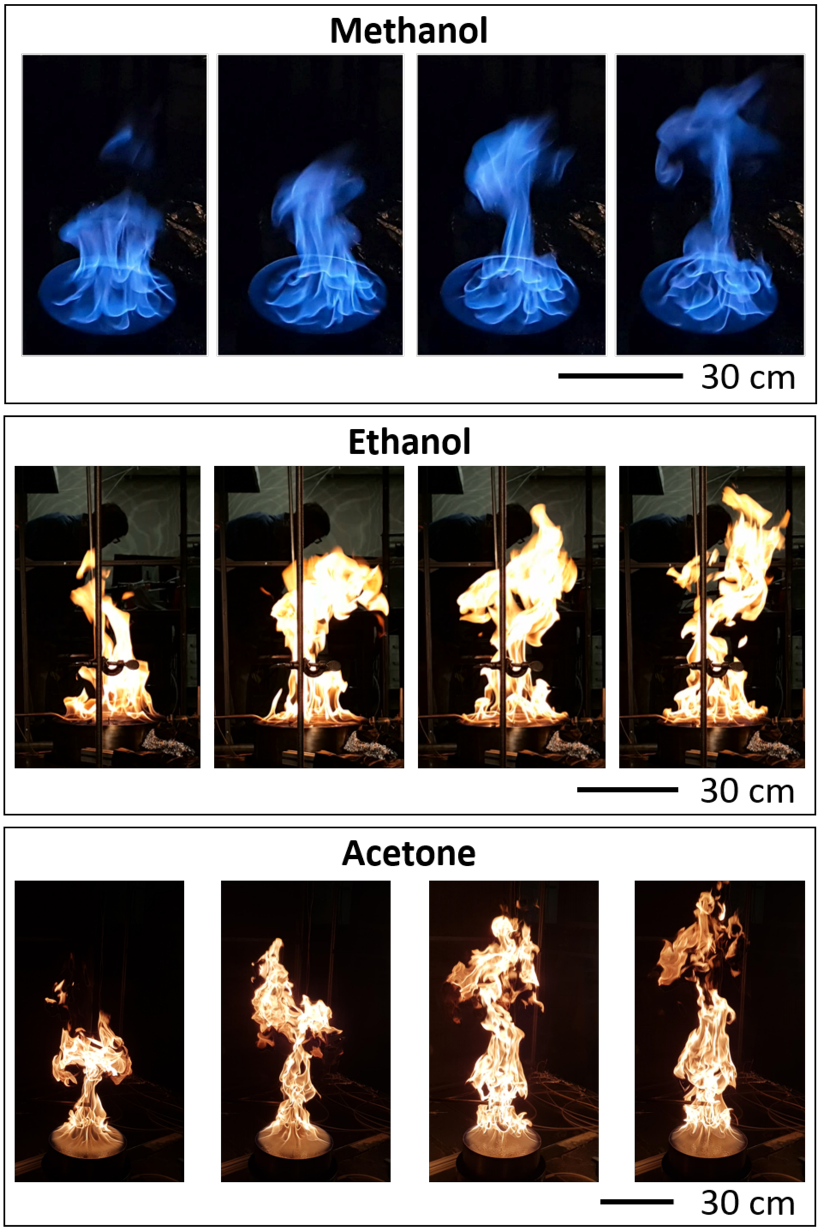


Fig. 3. Snapshots of the methanol (top), ethanol (middle), and acetone (bottom) pool fires during their pulsing cycles of frequency 2.6 Hz [9].

Table 1: List of measurements and thermochemical properties of fuels burning in a well-ventilated round 30 cm diameter pool fire burning in a quiescent environment. Measurement uncertainties are discussed in detail in Ref. [6].

Parameter (units)	**Methanol**	**Ethanol**	**Acetone**
Mass Burning Flux (g/m^2s)	12.4 ± 1.1	13.9 ± 0.8	17.6 ± 2.7
Heat Release Rate (kW)	17.4 ± 1.4	26.3 ± 1.5	35.5 ± 5.4
Mean Flame Height (cm)	36.4 ± 16.0	61.1 ± 28.2	91.5 ± 34.6
Heat of Combustion (kJ/g) [7]	19.9	26.8	28.6
C/H	3.0	4.0	6.0

The measured time-averaged burning flux and calculated ideal heat release rate are provided in Table 1. The heat release rate was calculated from Eq. 1. The time-averaged flame height reported in Table 1, of the methanol pool fire, was the shortest, followed by the ethanol, and then the acetone pool fires. In comparison to the measured mean flame height of each fuel, Heskestad's theoretical flame height, reported in Ref. [16], falls within the experimental uncertainty. The measurements in Table 1 are in agreement with measurements from Ref. [4] within experimental uncertainty.

3.2 Temperature and Gas Species Measurements of Methanol, Ethanol, and Acetone Pool Fires

To account for the difference in height between the methanol, ethanol, and acetone pool fires, the mean flame height, H, of each fuel was converted to dimensionless distance, Z^*, which was calculated as follows:

$$z^* = H\left(\frac{\dot{Q}}{c_p\sqrt{g}\rho_o T_o}\right)^{\frac{2}{5}} \tag{8}$$

Here $\dot{Q}$ is the heat release rate, g is the gravitational constant, and c_p and ρ_o are the specific heat and the density of air at room temperature, T_o.

Figure 4 shows the time-averaged corrected gas temperatures from the centerline of the methanol, ethanol, and acetone pool fires. The maximum mean temperature from each pool fire was found to peak at an approximate z^* of 0.6. Methanol was determined to have the highest mean temperature of 1316 K with ethanol and acetone exhibiting maximum mean temperatures of 1281 K and 1190 K, respectively. The maximum mean temperature of fuels with higher heat release rates is shown to be lower compared to other fuels.

Figure 5 displays the volume fraction, $\bar{X}_i$, of major species centerline measurements as a function

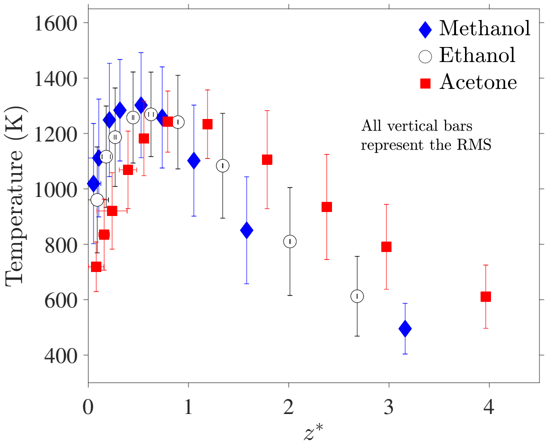


Fig. 4. Mean and RMS centerline temperature profiles of methanol, ethanol, and acetone pool fires during their pulsing cycles with z^* expanded uncertainty.

of Z^* along the centerline for the methanol, ethanol, and acetone fires. Major species detected in the TCD chromatogram include combustion reactants (fuels and oxygen, O_2), combustion products such as water, H_2O, and carbon dioxide, CO_2, combustion intermediates such as carbon monoxide, CO, hydrogen, H_2, and inert gases such as Nitrogen, N_2, and Argon, Ar. Methane was detected and quantified in all fires. In the case of the ethanol and acetone fires, soot, benzene, acetylene, ethylene, and ethane were also detected and quantified. Trace amounts of other species were also detected including propene, acetaldehyde, and ethyl acetate, which is consistent with the literature [17, 18].

For all fuels, the fuel and oxygen volume fractions were largest and smallest, respectively, close to the fuel surface. The volume fraction of fuel did not reach 1.0, since the lowest position of gas species measurement for all fuels was 2 cm above the burner rim. The volume fractions of inert gases were found to have increased relative to the distance from the fuel surface. Additionally, the maximum concentration of each species was observed at a lower z^* compared to where the maximum temperature was achieved.

Methanol and ethanol volume fraction profiles of H_2O and CO_2 with respect to Z^* were similar. The largest volume fractions for all product species, such as CO_2, CO, and H_2O, were achieved when burning acetone, except for H_2, whose peak was the lowest compared to the other fuels. Acetone was also found to produce a larger mass fraction of soot at every location along the centerline profile compared to the ethanol fire. No soot was observed in the methanol fire.

3.3 Comparison of Gas Species Measurements to Theoretical Values

Verification schemes were developed in order to assess the accuracy of the gas species measurements, carbon to hydrogen, product species, and inert species ratios were calculated and compared to theoretical values for the methanol, ethanol, and acetone pool fires. Each calculation

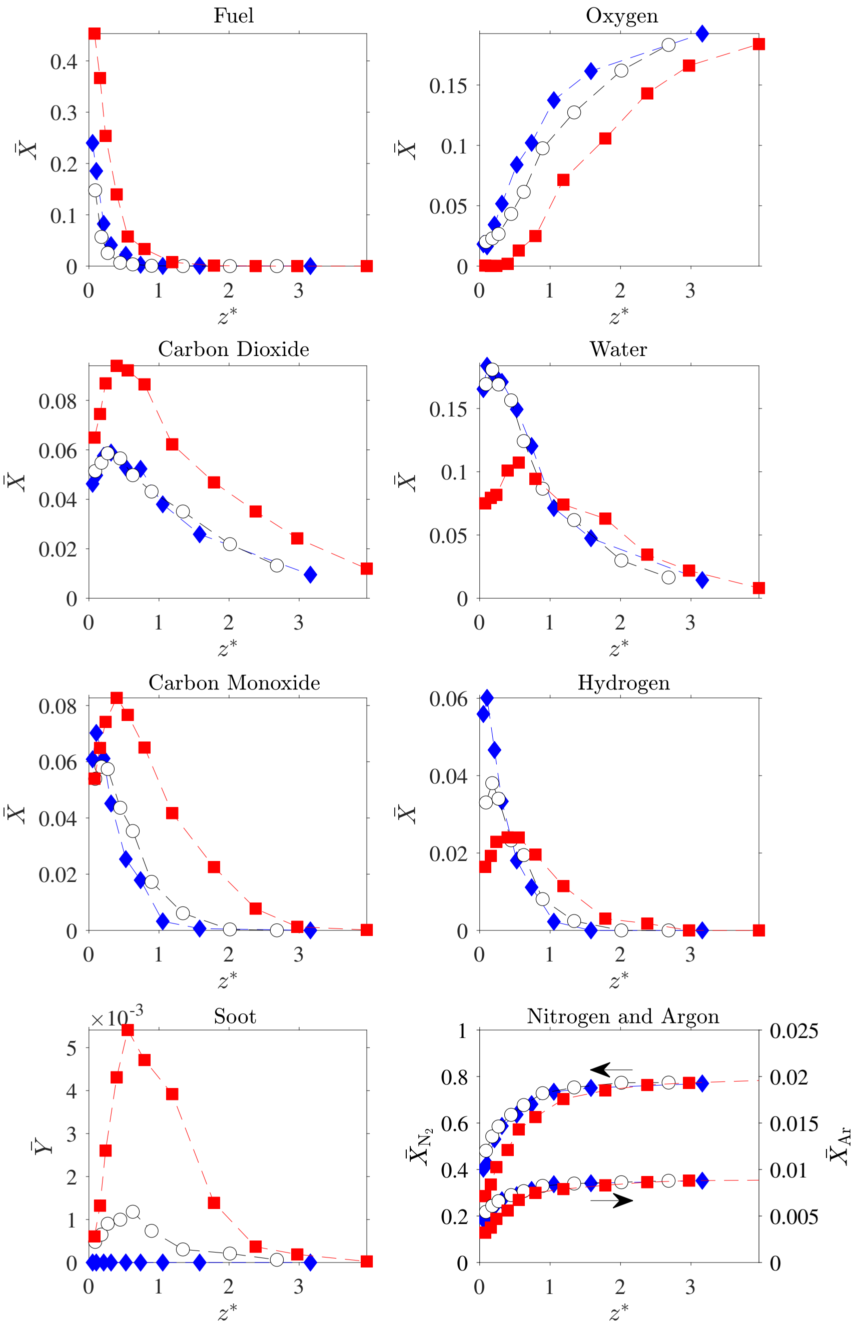


Fig. 5. Centerline volume fraction and soot mass fraction profiles of methanol (♦), ethanol (○), and acetone (■) pool fires.

only incorporates quantified species and assumes other identified species are trace amounts. The expanded uncertainty of the carbon to hydrogen, product, and inert ratios was determined using the law of propagation of uncertainty that accounted for the error of each volume fraction measurement.

The theoretical value of the carbon to hydrogen ratio was determined from the mass fraction of carbon and hydrogen residing within the parent fuel and are reported in Table 1. As shown in Eq. 9, the carbon to hydrogen ratio was calculated from the mass fraction of any quantified gas species that contained either carbon, $\bar{Y}_{i,\mathrm{C}}$, or hydrogen, $\bar{Y}_{i,\mathrm{H}}$. Here W_C and W_H are the molecular weights of carbon and hydrogen.

$$\frac{\mathrm{C}}{\mathrm{H}} = \frac{\sum \bar{Y}_{i,\mathrm{C}} \frac{W_\mathrm{C}}{W_i}}{\sum \bar{Y}_{i,\mathrm{H_2}} \frac{W_{\mathrm{H_2}}}{W_i}} \tag{9}$$

The carbon to hydrogen ratio for each experiment is shown in Fig. 6. The dotted line represents the theoretical value stated in Table 1. For each fuel, the data are shown to be in agreement with the theoretical value, indicating that the analysis is successful in quantifying most of the carbon and hydrogen containing species.

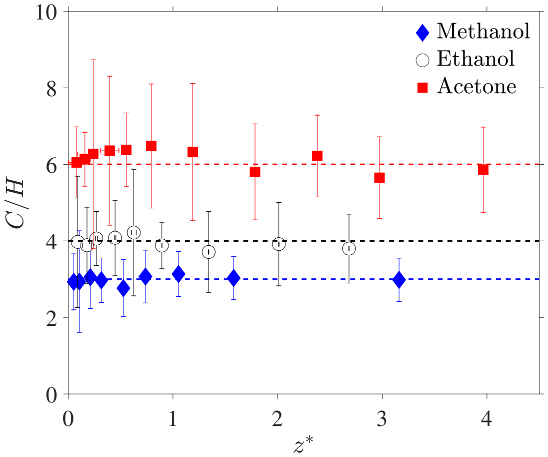


Fig. 6. Carbon to hydrogen ratio calculated from experimental values, with uncertainty bars, compared to theoretical values. Dotted lines represent the theoretical carbon to hydrogen ratio calculated from the mass fraction of carbon and hydrogen residing within the parent fuel.

Another verification test compared volume fraction measurements to stoichiometric combustion ratios, SCR, determined from the reaction below.

$$\begin{aligned} &\mathrm{C_xH_yO_z + a\ (O_2 + 3.76\ N_2 + 0.0445\ Ar)} \\ &\quad \mathrm{\rightarrow b\ C_xH_yO_z + c\ O_2 + d\ CO_2 + e\ H_2O + f\ H_2 + g\ CO} \\ &\quad\quad \mathrm{+ h\ CH_4 + i\ C_2H_2 + j\ C_2H_4 + k\ C_2H_6 + l\ C_6H_6} \\ &\quad\quad \mathrm{+ f\ (3.76\ N_2 + 0.0445\ Ar)} \end{aligned} \tag{10}$$

When simplified, the SCR of each fuel was calculated as:

$$\mathrm{SCR}_{\mathrm{methanol}} = \frac{\bar{X}_{\mathrm{H_2O}} + \bar{X}_{\mathrm{H_2}}}{\bar{X}_{\mathrm{CO_2}} + \bar{X}_{\mathrm{CO}}} = 2 \tag{11}$$

$$\mathrm{SCR}_{\mathrm{ethanol}} = \frac{\bar{X}_{\mathrm{H_2O}} + \bar{X}_{\mathrm{H_2}} + \frac{1}{2}\bar{X}_{\mathrm{CH_4}}}{\bar{X}_{\mathrm{CO_2}} + \bar{X}_{\mathrm{CO}} + \frac{2}{3}\bar{X}_{\mathrm{C_2H_4}} + 4\bar{X}_{\mathrm{C_6H_6}}} = \frac{3}{2} \tag{12}$$

$$\mathrm{SCR}_{\mathrm{acetone}} = \frac{\bar{X}_{\mathrm{H_2O}} + \bar{X}_{\mathrm{H_2}} + \bar{X}_{\mathrm{CH_4}} + \bar{X}_{\mathrm{C_2H_6}}}{\bar{X}_{\mathrm{CO_2}} + \bar{X}_{\mathrm{CO}} + 3\bar{X}_{\mathrm{C_6H_6}}} = 1 \tag{13}$$

Figure 7 shows the difference between the SCR and experimental data. The dotted lines represents the SCR values calculated from Eq. 11, 12, and 13. Direct comparison of the measured values to the idealized values shows that there is general agreement and that the volume fraction measurements are not unreasonable. An exact match is not expected since the idealized values do not take into account molecular diffusion, nor the mass of soot.

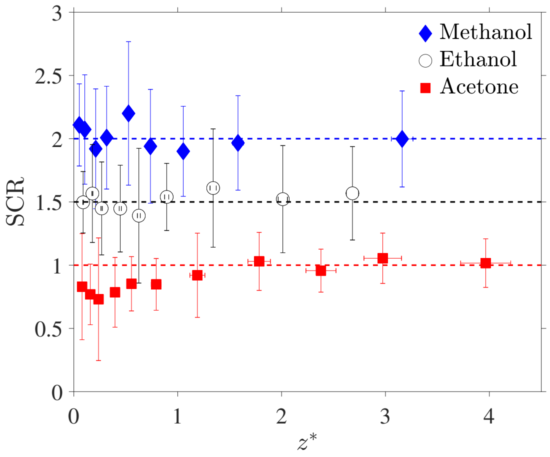


Fig. 7. Comparison of Stoichiometric Combustion Ratio, SCR, calculated from experimental and theoretical values. Dotted lines represent the theoretical SCR calculated using Eq. 12, 11, 13.

An inert ratio was also calculated from the volume fraction measurements of argon to nitrogen. Since both argon and nitrogen are inert, the ratio between them should be reasonably consistent across all fuels. The inert ratio was measured from ambient air samples using the setup described in Sec. 2.3. The inert ratio of ambient air was determined to be 0.012 ± 4 %. The range of the inert ratios calculated from the ambient air sample is depicted in Fig. 8 as the error band. The inert ratios determined from pool fire gas samples are shown to be within the error band region, which further supports the validity of the nitrogen and argon volume fraction measurements.

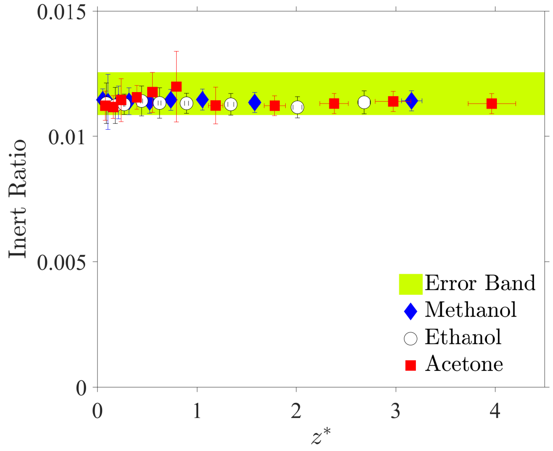


Fig. 8. Inert ratios calculated from the volume fractions for argon and nitrogen, compared to inert ratios determined from ambient air and represented as error band.

4. Conclusions

In summary, time-averaged local measurements of temperature and gas species concentrations were made to characterize the structure of methanol, ethanol, and acetone 30 cm diameter pool fire steadily burning in a quiescent environment. A verification scheme was developed to verify the gas species measurements that considered the overall stoichiometry of combustion for each fuel (see Eq. 12, 11, 13) Using this scheme, the gas species measurements were favorably compared to the idealized SCR values, which lends confidence to the veracity of the measurements. These local measurements complement previous measurements and provide insight into the complex chemical structure of medium-scale pool fires.

5. Acknowledgments

The authors would like to acknowledge Kimberly Harris of the Gas Sensing Metrology Group at NIST, who assisted in developing the species calibration method used in the study.

6. References

[1] K. McGrattan, S. Hostikka, R. McDermott, J. Floyd, C. Weinschenk, and K. Overholt, “Fire Dynamics Simulator, Technical Reference Guide” National Institute of Standards and Technology, Gaithersburg, Maryland, USA, and VTT Technical Research Centre of Finland, Espoo, Finland 6th edition, September 2013.

[2] S.J. Fischer, B. Hardouin-Duparc, and W.L. Grosshandler, “The structure and radiation of an ethanol pool fire,” Combustion and Flame vol. 70, no. 3 pp. 291–306, 1987.

[3] A. Hamins and A. Lock, “The Structure of a Moderate-Scale Methanol Pool Fire,” NIST Technical Note 1928 US Department of Commerce, National Institute of Standards and Technology, 2016.

[4] S.C. Kim, K. Y. Lee, and A. Hamins, “Energy balance in medium-scale methanol, ehtanol, and acetone pool fires,” Fire Safety Journal vol. 107, pp. 44–53, 2019.

[5] E.J. Weckman and A.B. Strong, “Experimental investigation of the turbulence structure of medium-scale methanol pool fires,” Combustion and Flame vol. 105, pp. 245–266, 1996.

[6] R. Falkenstein-Smith, K. Sung, J. Chen, and A. Hamins, “Mapping the chemical structure of centerline profiles in medium-scale pool fires,” NIST Technical Note, in preparation, US Department of Commerce, National Institute of Standards and Technology, submitted.

[7] DIPPR® Constant Properties for: Methanol, Ethanol, and Acetone https://dippr.aiche.org/SampleDb, (accessed 28 June 2019)

[8] J. Chen, Y. Zhao, X. Chen, C. Li, and C. Lu “Effect of Pressure on the Heat Transfer and Flame Characteristics of Small-Scale Ethanol Pool Fires” Fire Safety Journal vol. 99, pp. 27-37, 2018.

[9] Z. Wang, W.C. Tam, K.Y. Lee, J. Chen, and A. Hamins “Thin Filament Pyrometry Field Measurements in a Medium-Scale Pool Fire” Fire Technology accepted for publication, 2019.

[10] C.R. Shaddix, “A New Method to Compute the Radiant Correction of Bare-Wire Thermocouples,” 33rd ASME National Heat Transfer ConferenceTenth Mediterranean Combustion Symposium (MCS-10), Naples, Italy September 2017.

[11] F.P. Incropera, A.S. Lavine, T.L. Bergman, and D.P. DeWitt, Fundamentals of heat and mass transfer, 6th Edition, Wiley, 2007.

[12] A. Hamins, S. Fischer, T. Kashiwagi, M. Klassen, and J. Gore, “Heat feedback to the fuel surface in pool fires” Combustion Science and Technology vol. 97, no. 3, pp. 37–62, 1994.

[13] A. Hamins, M. Klassen, J. Gore, and T. Kashiwagi, “Estimate of flame radiance via a single location measurement in liquid pool fires” Combustion and Flame vol. 86, no. 3, pp. 223–228, 1991.

[14] A. Lock, M. Bundy, E. Johnson, A. Hamins, G. Ko, C. Hwang, P. Fuss, and R. Harris “Experimental study of the effects of fuel type, fuel distributution, and vent size on full-scale underventilated compartment fires in an iso9705 room,” NIST Technical Note 1603 US Department of Commerce, National Institute of Standards and Technology, 2008.

[15] C.D.A. Hogben, C.N. Young, E.J. Weckman, and A.B. Strong “Radiative properties of acetone pool fires,” 33rd ASME National Heat Transfer Conference, Albuquerque, NM August 1999.

[16] G. Heskestad “Luminous heights of turbulent diffusion flames” Fire safety Journal vol. 5, pp.103-108, 1983.

[17] S. Pichon, G. Black, N. Chaumeix, M. Yahyaoui, J.M. Simmie, H.J. Curran, and R. Donohue “The combustion chemistry of a fuel tracer: Measured flame speeds and ignitiondelays and a detailed chemical kinetic model for the oxidation of acetone” Combustion and Flame vol. 156, pp.494-504, 2009.

[18] J. Gong, S. Zhang, Y. Cheng, Z. Huang, C. Tang, and J. Zhang “A comparative study of n-propanol, propanal, acetone, and propane combustion in laminar flames.” Proceedings of the Combustion Institute vol. 35, pp.795-801, 2015.

Figure captions

Fig. 1. 30 cm pool burner with fuel level indicator, overflow section, and water-cooled gas sampling probe.

Fig. 2. A schematic of the extractive sampling setup used to extract and transport fire samples from the pool fire to the GC/MS.

Fig. 3. Snapshots of the methanol (top), ethanol (middle), and acetone (bottom) pool fires during their pulsing cycles of frequency 2.6 Hz [9].

Fig. 4.Mean and RMS centerline temperature profiles of methanol, ethanol, and acetone pool fires during their pulsing cycles with z^* expanded uncertainty.

Fig. 5. Centerline volume fraction and soot mass fraction profiles of methanol (⧫), ethanol (◯), and acetone (■) pool fires.

Fig. 6. Carbon to hydrogen ratio calculated from experimental values, with uncertainty bars, compared to theoretical values. Dotted lines represent the theoretical carbon to hydrogen ratio calculated from the mass fraction of carbon and hydrogen residing within the parent fuel.

Fig. 7. Comparison of Stoichiometric Combustion Ratio, SCR, calculated from experimental and theoretical values. Dotted lines represent the theoretical SCR calculated using Eq. 12, 11, 13.

Fig. 8. Inert ratios calculated from the volume fractions for argon and nitrogen, compared to inert ratios determined from ambient air and represented as error band.

40th AIVC - 8th TightVent & 6th venticool Conference

The Role of Carbon Dioxide in Ventilation and IAQ Evaluation: 40 years of AIVC

Andrew Persily
National Institute of Standards and Technology
100 Bureau Drive, MS8600
Gaithersburg, MD 20899 USA
**Corresponding author: andyp@nist.gov*

SUMMARY

The purpose of this summary is to review Air Infiltration and Ventilation Centre activities, as reflected in its publications, related to indoor carbon dioxide over the 40 years that have transpired since its creation. These activities, like most applications of indoor CO_2 to the fields of ventilation and indoor air quality, have focused on the following: control of outdoor ventilation rates, i.e., demand control ventilation; use as a tracer gas to measure outdoor air change rates; providing an indicator or metric of IAQ; and, directly impacting human health, comfort and performance. More recent work on CO_2 generation rates from building occupants and CO_2 concentrations in standards and building regulations is also covered. This summary was generated by searching on Air Infiltration and Ventilation Centre publications, though the findings also reflect the evolving application and understanding of indoor CO_2 in the broader literature.

KEYWORDS: carbon dioxide; indoor air quality; metrics; ventilation

GENERAL DISCUSSIONS OF INDOOR CO_2

One of the earliest Air Infiltration and Ventilation Centre (AIVC) publications on the application of indoor CO_2 is a short article covering a range of topics, including tracer gas applications, indoor air quality (IAQ) evaluation, and CO_2 as an indicator of occupancy (Liddament, 1996). Another short paper was published more recently, which focused on CO_2 as an IAQ indicator and for ventilation control (de Gids and Wouters, 2010). No other general reports or publications on CO_2 have been issued by the AIVC over its 40 years. The application of CO_2 has been covered mostly by individual conference papers on the topics covered below. For each of the topics, a table of references is provided at the end of this summary. These tables are not exhaustive but provide a sense of the issues covered for each topic.

DEMAND CONTROL VENTILATION

Indoor CO_2 has been discussed as a control parameter for outdoor air ventilation for decades, with the goal being to provide sufficient ventilation for the occupants in a space. Ventilating for the actual occupancy rather than a maximum design value provides an opportunity to reduce energy used for space heating and cooling, as well as assuring that the ventilation is sufficient to meet the needs of the occupants. In 2001 the AIVC generated a literature list (LL) that identified about 50 publications on the topic of CO_2 demand control ventilation, many of them not published by the AIVC itself. Additional work on the topic has continued in subsequent publications on sensor performance, energy and IAQ impacts, case studies in a variety of building types and other subtopics as noted in the table below.

CO_2 AS A TRACER GAS

Carbon dioxide has long been recognized as a useful tracer gas for studying building ventilation and airflow given its low reactivity and toxicity, relative ease of measurement and, in some

Persily, Andrew K.; Dols, William Stuart; Emmerich, Steven; Ng, Lisa. "The Role of Carbon Dioxide in Ventilation and IAQ Evaluation: 40 years of AIVC." Paper presented at 40th Air Infiltration and Ventilation Centre Conference, Ghent, BE. October 08, 2019 - October 09, 2019.

applications, building occupants serving as a convenient tracer gas source. CO_2 was identified as a potential tracer gas in an early AIC publication (Liddament and Thompson, 1983). Since that time, CO_2 has been used as a tracer gas in many studies, with several noted below.

IAQ ASSESSMENT

Indoor CO_2 concentrations have long been used as part of IAQ assessments with the oldest reference listed in the table below dating back to 1985. Some of these assessments measure CO_2 concentrations as one of many pollutants monitored, though many assessments do not explain the significance of the measured concentrations or compare them to a reference or guideline value. Such measurements are still common as part of IAQ investigations; the explicit consideration of CO_2 concentration metrics is a more recent development and is discussed next.

IAQ METRIC

The AIVC has focused on IAQ metrics in recent years, with the topic being a major theme of its 2016 conference held in conjunction with the ASHRAE IAQ conference series. Only two papers on the topic of CO_2 as an IAQ metric are listed in the table below, but the issue has been discussed in recent AIVC workshops and conference sessions without any papers being published and those discussions are likely to continue.

CO_2 GENERATION RATES

The use of CO_2 as a tracer gas for quantifying building and space ventilation rates requires a value of the rate of CO_2 generation by the building occupants. For many years, default values from ASHRAE and other sources have been used without evaluating their accuracy or the sources on which they were based. Recent publications have developed more well-documented and robust methods for estimating these generations rates, with three AIVC conference papers included in the table below.

STANDARDS AND REGULATIONS

While indoor CO_2 has been considered in ventilation and IAQ studies for decades, most standard or guideline values were only for industrial environments. More recently a number of standards and building regulations have been promulgated with specific indoor CO_2 concentration limits. Several of these are covered by the publications listed in the table below, though other countries and localities appear to also be setting such limits.

CO_2 IMPACTS ON BUILDING OCCUPANTS

Finally, a number of recent studies have taken a new look at how CO_2 impacts building occupants both physically and mentally. Many of these studies have been looking at concentrations that are typical of indoor spaces. However, the studies in the broader literature are not consistent as to the human effects observed. The three studies listed in the table below are just an example of such work that has been presented in recent AIVC conferences.

REFERENCES

de Gids, WF and Wouters, P. (2010). *CO_2 as Indicator for the Indoor Air Quality - General Principles*, Air Infiltration and Ventilation Centre.

Liddament, M and Thompson, C. (1983). *Techniques and Instrumentatoin for the Measurement of Air Infiltration in Buildings - A Brief Review and Annotated Bibliography*, Air Infiltration Centre, Bracknell, Great Britain., Technical Note 10.

Liddament, MW. (1996). Why CO_2? *Air Infiltration Review*, 18, 1-4.

Demand Control Ventilation
H. Han, K-J Jang, C. Han and J. Lee. 2013. Occupancy estimation based on CO2 concentration using dynamic neural network model, 34th AIVC Conference.
A. Persily, A. Musser, S. Emmerich, M. Taylor. 2003. Simulations of indoor air quality and ventilation impacts of demand controlled ventilation in commercial and institutional buildings. 24th AIVC and BETEC Conference.
Villenave J.G., Bernard A.M., Lemaire M.C. 2003. Simulations of indoor air quality and ventilation impacts of demand controlled ventilation in commercial and institutional buildings. 24th AIVC and BETEC Conference.
Chan G Y, Chao C Y, Lee D C, Chan S W, Lau H. 1999. Development of a demand control strategy in buildings using radon and carbon dioxide levels. Indoor Air 99 and 20th AIVC Conference.
Fleury B. 1992. Demand controlled ventilation: a case study. 13th AIVC Conference.
Zamboni M, Berchtold O, Filleux C, Fehlmann J, Drangsholt F. 1991Demand controlled ventilation - an application to auditoria. 12th AIVC Conference.
Fahlen P, Andersson H. 1991. Demand controlled ventilation: full scale tests in a conference room. 12th AIVC Conference.
Donnini G, Haghighat F, Van Hiep Nguyen. 1991. Ventilation control of IAQ, thermal comfort and energy conservation by CO2 measurement. 12th AIVC Conference.
Fahlen P, Ruud S, Andersson H. 1991. Demand controlled ventilation - evaluation of commercially available sensors. 12th AIVC Conference.
Raatschen W. 1988. Market analysis of sensors for the use in demand controlled ventilating systems. 9th AIVC Conference.
Smith B E, Prowse R W, Owen C J. 1984. Development of occupancy-related ventilation control for Brunel University Library. 5th AIVC Conference.

Use of CO_2 as a Tracer Gas
J.D. Carrilho, M. Mateus, S. Batterman, M. Gameiro da Silva. 2014. Measurement of infiltration rates from daily cycle of ambient CO_2. 35th AIVC Conference.
D. Kraniotis, T. Aurlien, T.K. Thiis. 2013. On investigating instantaneous wind-driven infiltration rates using CO2 decay method. 35th AIVC Conference.
Bong C, Kim S, Lee J, Lee H. 1999. Ventilation demand in a subway train - based on CO2 bioeffluent from passengers. Indoor Air 99 and 20th AIVC Conference.
Federspiel C. 1996. Ventilation performance evaluation using passively-generated carbon dioxide as a tracer gas. 17th AIVC Conference.
Ekberg L E, Strindehag O. 1996. Checking of ventilation rates by CO2 monitoring. 17th AIVC Conference.
Kohal J S, Riffat S B, 1993. Computer modelling & measurement of airflow in an environmental chamber. 14th AIVC Conference.

IAQ Assessment
J. Sifnaios, P.V.Dorizas, M. Assimakopoulos. 2014. A study of carbon dioxide concentrations in elementary schools. 35th AIVC Conference.
Weinlader H, Beck A, Fricke J, 2000. Demand controlled ventilation in schools - energetic and hygienic aspects. 21st AIVC Conference.
Parent D, Stricker S, Fugler D, 1996. Ventilation in houses with distributed heating systems. 17th AIVC Conference.
Donnini G, Nguyen V H, Molina J, 1994. Occupant satisfaction and ventilation strategy - a case study of 20 public buildings. 15th AIVC Conference.
Weinlader H, Beck A, Fricke J, 2000. Demand controlled ventilation in schools - energetic and hygienic aspects. 21st AIVC Conference.
Parent D, Stricker S, Fugler D, 1996. Ventilation in houses with distributed heating systems. 17th AIVC Conference.
Donnini G, Nguyen V H, Molina J, 1994. Occupant satisfaction and ventilation strategy - a case study of 20 public buildings. 15th AIVC Conference.
Grelat A, Cohas M, Lemaire M C, Fauconnier R, Creuzevault D, Loewenstein J-C, 1992. Correlation between carbon dioxide concentration and condensation in homes. 13th AIVC Conference.
Nielsen J B, 1992. A new ventilation strategy for humidity control in dwellings. 13th AIVC Conference.
Croome D J, Gan G, Awbi H B, 1992. Field evaluation of the indoor environment of naturally ventilated offices. 13th AIVC Conference.

Fehlmann J, Wanner H U, 1990. Air change rate and indoor air quality in bedrooms of well tightened residential buildings. 11th AIVC Conference.
Grot R A, Persily A, Hodgson A T, Daisey J M, 1988. Ventilation and indoor air quality in a modern office building. 9th AIVC Conference.
Baumgartner T, Bruhwiler D, 1987. Simulation of CO2 concentration for determining air change rate. 8th AIVC Conference.
Fecker I, Wanner H U, 1986. Measurement of carbon dioxide of the indoor air to control the fresh air supply. 7th AIVC Conference.
Lundqvist G R, 1985. Indoor air quality and air exchange in bedrooms. 6th AIVC Conference.

IAQ Metric
A. Persily. 2018. Development of an Indoor Carbon dioxide metric. 39th AIVC Conference.
A. Szczurek, M. Maciejewska, T. Pietrucha. 2015. CO2 and volatile organic compounds as indicators of IAQ. 36th AIVC Conference.

CO_2 Generation Rates
M. Tajima, T. Yorimitsu, Y. Shimada. 2018. Accuracy Improvement for Estimating Indoor Carbon Dioxide Concentration Produced by Occupants. 39th AIVC Conference.
A. Persily, L. de Jonge, 2017. A New Approach to Estimating Carbon Dioxide Generation Rates from Building Occupants. 38th AIVC Conference.
M. Tajima, T. Inoue, Y. Ohnishi. 2014. Derivation of equation for personal carbon dioxide in exhaled breath intended to estimation of building ventilation. 35th AIVC Conference.

Standards and Regulations
S. Caillou, J. Laverge, P. Wouters. 2018. IAQ in working environments in Belgium: alternative approaches to CO2 requirement. 39th AIVC Conference.
A. Persily. 2015.Indoor Carbon Dioxide Concentrations in Ventilation and Indoor Air Quality Standards. 36th AIVC Conference.
P. Paulino. 2015. Impact of the new rite 2013 (regulation on thermal installation) on indoor air quality. 36th AIVC Conference.

CO_2 Impacts on Building Occupants
L. Yoshimoto, T. Yamanaka, A. Takemura, K. Ikeda. 2018. Subjective Evaluation for Perceived Air Pollution Caused by Human Bioeffluents. 39th AIVC Conference.
P. Wargocki, J.A. Porras-Salazar, W.P. Bahnfleth, 2017. Quantitative relationships between classroom CO2 concentration and learning in elementary schools. 38th AIVC Conference.
X. Zhang, P. Wargocki, Z. Lian, 2015. Effects of Carbon Dioxide With and Without Bioeffluents on humans. 36th AIVC Conference.

Prevention of Cooktop Ignition Using Detection and Multi-Step Machine Learning Algorithms

Wai Cheong Tam[a*], Eugene Yujun Fu[b*], Amy Mensch[a], Anthony Hamins[a*], Christina You[c], Grace Ngai[b], Hong va Leong[b]

[a]National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD, USA, {waicheong.tam, amy.mensch, anthony.hamins}@nist.gov

[b]The Hong Kong Polytechnic University, 11 Yuk Choi Road, Kowloon, Hong Kong, China, {csyfu, csgngai, cshleong}@comp.polyu.edu.hk

[c]Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PE, USA, cyou2@andrew.cmu.edu

*Corresponding author

Highlights:

- Time series sensor data for cooking with electric-coil and gas cooktops are presented.
- Discussion of data preprocessing and feature selection for machine learning on real-time fire prevention application.
- Workflow for constructing and testing machine learning models is demonstrated.
- Support vector machine (SVM) detection algorithm correctly classifies 96.9 % of the data points as normal or hazardous conditions for oils on an electric-coil cooktop.
- Multi-step models can enhance data classification and improve overall detection accuracy.

Abstract:

This paper[1] presents a study to examine the potential use of machine learning models to build a real-time detection algorithm for prevention of kitchen cooktop fires. Sixteen sets of time-dependent sensor signals were obtained from 60 normal or ignition cooking experiments. More than 200 000 data instances are documented and analyzed. The raw data are preprocessed. Selected features are generated for time series data focusing on real-time detection applications. Utilizing the leave-one-out cross validation method, three machine learning models are built and tested. Parametric studies are carried out to understand the diversity, volume, and tendency of the data. Given the current dataset, the detection algorithm based on support vector machine (SVM) provides the most reliable prediction of pre-ignition conditions (with an overall classification accuracy of 96.9 % for oils on an electric-coil cooktop). Analyses indicate that using a multi-step approach could further improve overall prediction accuracy. The development of an accurate detection algorithm can provide reliable feedback to intercept ignition of unattended cooking and help reduce fire losses.

Keywords: machine learning, time series classification, cooking, fire prevention, fire detection

[1] Certain commercial products are identified in this paper in order to specify adequately the equipment used. Such identification does not imply recommendation by the National Institute of Standards and Technology, nor does it imply that this equipment is the best available for the purpose.

Tam, Wai Cheong; Fu, Eugene Yujun; Mensch, Amy; Hamins, Anthony; You, Christina; Ngai, Grace; Leong, Hong va. "Prevention of Cooktop Ignition Using Detection and Multi-Step Machine Learning Algorithms." Paper presented at International Association of Fire Safety Science 2020, Waterloo, CA. April 27, 2020 - May 01, 2020.

1. Introduction

A study conducted by the National Fire Protection Association (NFPA) based on fire loss data from the National Fire Incident Reporting System summarizes the US cooking fire problem [1]. During the period 2013 – 2017, household fires involving cooking were responsible for approximately 173 200 fires, 5 020 injuries, and 550 deaths annually. This represents 49 % of all reported home fires in the US. In 2017, U.S. Fire Departments responded to an average of 470 home cooking fires per day. Of these fires, cooktops (or ranges) were found to be involved in 62 % of reported home cooking fire incidents, 89 % of home cooking fire deaths, and 79 % of home cooking fire injuries. Unattended cooking was the leading cause of cooking fires.

Households that use electric ranges have a higher risk of developing a cooking fire than those using gas ranges. Although 60 % of households cook with electric cooktops, 79 % of reported cooktop fires were electric [1]. In terms of standards development, the abnormal ignition test in UL 858[2] represents significant progress in addressing unattended electric-coil heating element cooktop fires [2]. However, there are no current analogous standards applicable to other types of electric cooktops or to gas ranges in North America, and there are no standards for existing electric-coil element cooktops/ranges manufactured before 2019. Additionally, the cooking fire problem in the US may be worse than it was a few decades ago. Cooking caused more home fire deaths in 2013-2017 than in 1980-1984 [1]. In summary, the cooking fire problem has not been adequately addressed and the consequent fire casualties. New approaches are needed to reduce cooking fires.

Typical smoke detectors being installed in kitchens are well-known to be prone to nuisance alarms caused by normal cooking activities. Despite research efforts made to understand the cooking fire characteristics [3,4], the signal behavior of various sensors [5,6] and the development of advanced cooking ignition detection algorithms [7,8], only a few of these improvements can be used in practical environments. Indeed, studies [9,10] show that commercially available smoke detectors are primarily designed to provide alerts and/or warnings for flaming fires in which fuel packages such as upholstered furniture, appliances, and/or kitchenware are being ignited. For detection of cooktop pre-ignition conditions, alterative detection technologies and analysis are needed.

Johnsson [11] conducted a series of experiments to determine the feasibility of using research-grade sensors to distinguish the signal behaviors between normal cooking activities and ignition conditions on an electric range in a mock kitchen. Optimal location for sensor placement and the associated cut-off value for such a detection system were suggested. More recently, Mensch et al. [12] extended the work of Johnsson using a wide range of consumer-grade sensors to develop a pre-ignition detection system. Sensor signal differences were investigated, and a threshold based on the magnitude of a volatile organic compounds (VOC) sensor was determined to best differentiate the signal peak for normal cooking and the minimum signal 60 s before ignition. More recently, the pre-ignition detection system was extended to account for both normal and ignition cooking conditions using gas cooktops [13]. Mensch and co-workers conducted a statistical analysis and showed that the previously determined threshold for the VOC sensor had to be modified to accommodate the additional data and to assure detection sensitivity and

[2] An "abnormal" ignition test was extended to consider the average temperature of a dry (without cooking oil), round, 20 cm diameter, cast iron pan [2]. If the average pan temperature does not exceed 385 °C for 30 min with the element on its highest power setting, the test is considered a pass.

nuisance immunity. A machine learning algorithm was implemented to study the prediction performance of using signals from either individual sensor or all sensors. The preliminary results highlighted the potential benefits of implementation of machine learning algorithms. Indeed, as the volume of data becomes larger and data complexity increases, advanced data processing, such as considering the rate of change of a sensor signal or ratios of signals from different sensors, may be beneficial to discriminate different cooking conditions.

Machine learning algorithms have the capability to transform a large set of complex data with multiple variables in meaningful ways, such that a hyperplane or a dynamic threshold can be obtained for classification problems. In fact, data-driven analytics have been used to resolve complex problems for a variety of applications. In structural engineering [14], a logistic regression model was used to provide failure detection of a shackle with a dual sensing system. For outage prediction of power grids [15,16], three-dimensional support vector machine (SVM) [17] were developed using limited data. Hand-crafted features were extracted to improve the performance of prediction. In the combustion community [18], a convolutional neural network was used to predict the likelihood of ignition for a hydrogen jet in air crossflow. However, no previous studies have been carried out to study the use of machine learning models that can classify pre-ignition conditions for different cooking scenarios. Using the experimental data collected in [13], this paper contributes the use of machine learning for development of robust and reliable early warning algorithms for sensor data. The developed algorithm can help to automatically cut cooktop power or gas flow to prevent cooktop ignition.

The rest of this paper is divided into four sections. Section 2 presents the experimental apparatus and procedure. Section 3 presents detailed descriptions associated with the proposed machine learning models. Section 4 discusses the results and key findings. Conclusions are provided in Section 5.

2. Experimental Apparatus and Procedure

Here, the focus is on electric coil element cooktop fires although some experiments were conducted using a gas cooktop. The experimental apparatus and procedure have been previously described in [12,13,19]. Some details are included here to provide a general description of the work.

2.1 General Setup

A series of 60 experiments were conducted in a mock kitchen, previously reported in [12,13]. A schematic of the experimental setup is shown in Figure 1. As seen in the figure, a cooktop/range is located between two gypsum-board cabinets. The dimensions of the cooktop/range are 68 cm in depth and 76 cm in width. The top surface of the cooktop/range was level with the standing cabinets. Two cooktops were considered: a household electric-coil cooktop with two heating element sizes and a household gas cooktop with three heating element sizes. For the electric-coil stove, both the big (20 cm) and the small (15 cm) heating elements were used. The heating power was measured as approximately 1.1 kW and 1.8 kW for the 15 cm and 20 cm electric-coils, respectively. For the gas cooktop, the estimated heating power was 4.0 kW for the large burner and 3.4 kW for the medium burner. The small gas burner was not used in the experiments.

An aluminum shroud was used in some of the experiments to reduce the influence of air circulation in the room.

2.2 Range Hood and Duct

Figure 1 shows the mock kitchen with a nominal 76 cm wide exhaust hood. The separation distance between the bottom surface of the range hood and the upper surface of the cooktop was 84 cm. The range hood was a 200-CFM (approximately 0.1 m^3/s) rated venting system to remove smoke, grease, odors, and moisture from the cooking space. The fan had variable flow control. The outlet of the range hood was connected to a nominal 15 cm diameter aluminum duct and the exhaust air was vented to the outside environment. The exhaust flow was characterized using a velocity probe at the center of the duct after the sensor array. The average flow speed and its standard deviation was 3.4 m/s ± 0.1 m/s over all the experiments. Using the electric coil cooktop, the duct temperature increased by an average of 9 °C during cooking, causing an estimated reduction in duct mass flow of 3 %. For the gas cooktop, the duct temperature increased by an average of 23 °C, which is estimated to reduce the duct mass flow by 7 %. The duct length between the range hood opening and the sensors was approximately 3 m.

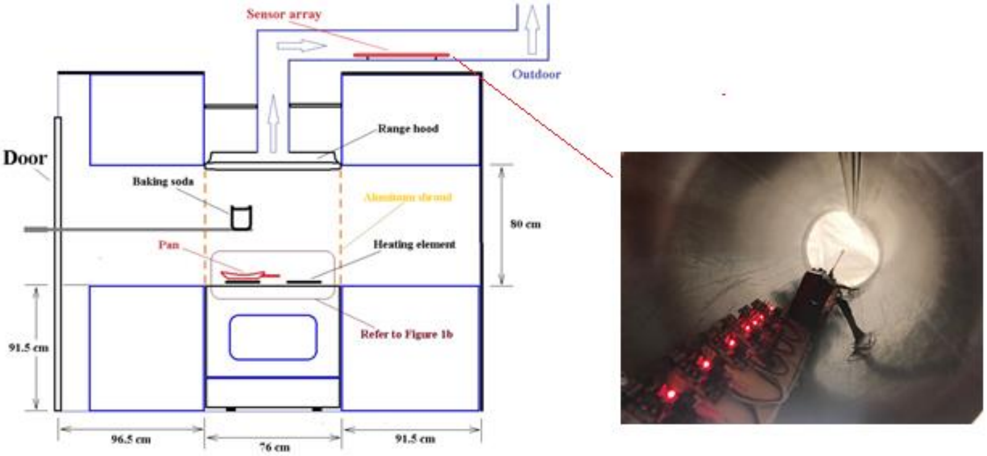


Figure 1. Schematic drawing of experimental setup (not to scale) and close-up view of the sensor array (in the duct).

2.3 Sensor Array

Figure 1 shows the sensor array secured inside the duct. There were 14 different sensors which are sensitive to smoke, alcohol, hydrocarbons, hydrogen, natural gas, carbon monoxide, volatile organic compounds (VOC), dust/aerosols, humidity, flow, indoor air quality (IAQ), temperature, and carbon dioxide. The dust sensor was modified from a commercial product to extend its range of sensitivity. Appendix A provides additional information for the sensor array.

2.4 Cooking Pans

A number of cooking pans were used in the experiments. The pans were either 20 cm or 25 cm diameter round pans. Four different types of pan materials were considered: cast iron, aluminum, stainless-steel, and a 5-layer aluminum/stainless-steel composite. Most of the experiments used the 20 cm cast iron pan, which is the cookware specified in the new UL 858 standard test [2]. It should also be noted that Type-K thermocouples were spot-welded or peened onto the top surface of each pan to monitor its temperature. In general, there were two thermocouples used to measure the pan temperature: one at the center of the pan, and one closer to the edge. Thermocouples were used to monitor the stage of cooking. For the 20 cm cast iron pan on the small electric burner, the temperature towards the edge of the pan was about 23 °C higher than the center temperature. For experiments that only measured the temperature in the center of the pan, the temperature closer to the edge was estimated using a linear regression of the relationship between the two temperatures from experiments that both temperatures were measured [13].

2.5 Food

A representative range of foods were considered, including cooking oil, butter, chicken legs, salmon, hamburger, bacon, and french fries. As mentioned in [12,13], the cooking oils tested included canola oil, corn oil, olive oil, sunflower oil, and soy oil. The most common experiment used 50 mL of oil, but larger volumes of oil were also considered. A list of important test conditions for all experiments is summarized in Appendix B.

2.6 Test Procedure and Data Acquisition

For most experiments, both oils and food samples were prepared and placed on the pan before turning on the burner. The pan was centered on the specified burner. The gas or electric burner was generally set to its maximum power setting, and the output from the sensors was monitored. Data acquisition started at least 180 s before the heating element was turned on. Experiments were conducted until the food samples ignited or charred, or until normal cooking was complete. If ignition occurred, the fire was extinguished immediately by remotely applying baking soda. The range hood was kept turned on for the entire test period. Background signal values for each sensor, obtained before the burners were turned on, were subtracted from the raw signals. LabVIEW was utilized and an in-house program was coded to facilitate data acquisition. The sampling rate was set to 0.25 Hz. In total, data was obtained across over 12 800 time points.

3. Procedure for Machine Learning

Depending on the nature of the data (i.e., time series, images, text), the optimal machine learning architecture may be different. However, the overall workflow is relatively the same. After data collection, there are five primary steps: data profiling, preprocessing, feature selection, training and evaluation.

3.1 Data profiling

Figure 2 shows the data for 11 selected sensors together with the pan temperature for Exp. 8, which tested 50 mL of canola oil in a round 20 cm cast iron pan on a small electric cooktop burner set to maximum power at time = 0. When the pan temperature is below 200 ºC, there is nearly no change to any of the sensor signals. At approximately 230 ºC, most of the signals begin to rise increase rather monotonically, which was generally true for all the oil experiments. The red dashed line indicates the cooking condition, either normal or pre-ignition. The curve changes from 0 to 1 when the pan temperature exceeds 300 ºC at about 320 s. At 704 s, the pan temperature reached 429 ºC, and auto-ignition occurred; the test was immediately terminated.

Tam, Wai Cheong; Fu, Eugene Yujun; Mensch, Amy; Hamins, Anthony; You, Christina; Ngai, Grace; Leong, Hong va. "Prevention of Cooktop Ignition Using Detection and Multi-Step Machine Learning Algorithms." Paper presented at International Association of Fire Safety Science 2020, Waterloo, CA. April 27, 2020 - May 01, 2020.

Figure 3 shows 3 sets of curves for Exp 8 (solid line)[3], Exp 46 (circle symbols)[4], and Exp 57 (triangle symbols)[5]. Each set of curves contain the sensor signals for IAQ (black) and VOC (green), as well as the red line denoting the end of normal cooking and the beginning of the pre-ignition condition. Sensor data were selected based on the clear independence of the sensor signals from each other. For the two oil experiments (Exp 8 and Exp 57), the IAQ signals monotonically increased with time. However, the profile of the IAQ signal associated with Exp 46 was observably different; it first increased to a value of 0.3 x 10^4, remained constant for more than 300 s, then increased to a peak value of 0.76 x 10^4, reached a minimum value of 0.4 x 10^4, before obtaining another peak value of 1.1 x 10^4. The VOC sensor for Exp 46 also shows the same complex behavior. The physical mechanism causing such behavior is not known. However, this test was repeated and the same behavior was observed. Sensor data associated with all experiments were examined; Exp 16 was eliminated because the pan was not cleaned before the experiment such that a layer of degraded oil was on the bottom of the pan before the fresh oil was added. In the future, this type of experiment could be considered.

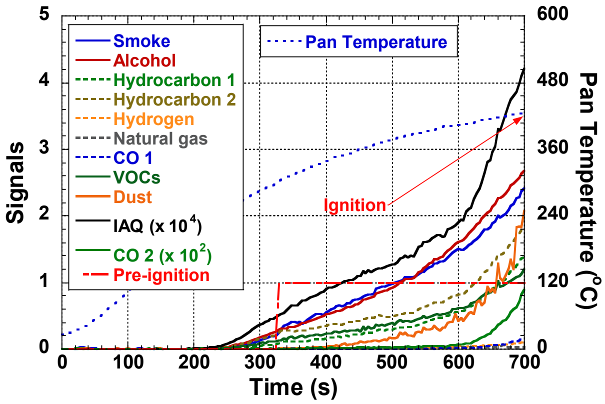


Figure 2. Signals for 11 selected sensors during Experiment 8.

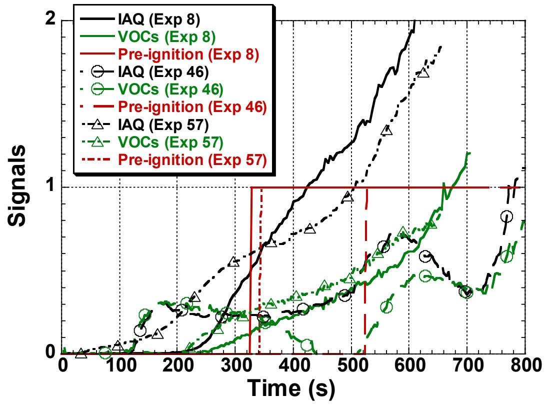


Figure 3. Comparison of IAQ (divided by 10^4) and VOC signals for 3 experiments (8, 46, 57).

[3] Exp 8 is for 50 mL canola oil on a 20 cm cast iron pan heated by the small electric-coil burner.

[4] Exp 46 is for 110 g bacon on a 20 cm cast iron pan heated by the small electric-coil burner.

[5] Exp 57 is for 50 mL canola oil on a 20 cm cast iron pan heated by the big gas burner.

3.2 Preprocessing

As illustrated in Fig 2, the data range significantly differed between some of the sensor signals. For instance, in Exp 8 the data range of the smoke signal was from 0 to 2.5 V, while the IAQ signal range was more than 48 000. The difference in magnitude associated with the data is known to affect training efficiency, so a min-max normalization was applied to all signals to impose a range from 0 to 1.

In machine learning, a classifier is applied to a set of data instances that are well-annotated. The classifier learns the data distribution, or fits a hyperplane in multi-dimensional space, to separate data instances based on feature vectors and corresponding labels. Here, data instances are generated using a moving window. The advantage of this approach is to provide more information for each data instance, such as temporal information for the historical data.

The analysis applies a moving time window of size, W, as demonstrated in Figure 4. Each moving window represents one data instance. The measurement frequency of each signal was f = 0.25 Hz (1 sample every 4 s) and there were W x f sample points for each signal in a moving window. Six signals are considered here, measuring alcohol, CO, dust, indoor air quality (IAQ), smoke, and VOC, leading to 6 time series with W x f sample points for each instance of processed data.

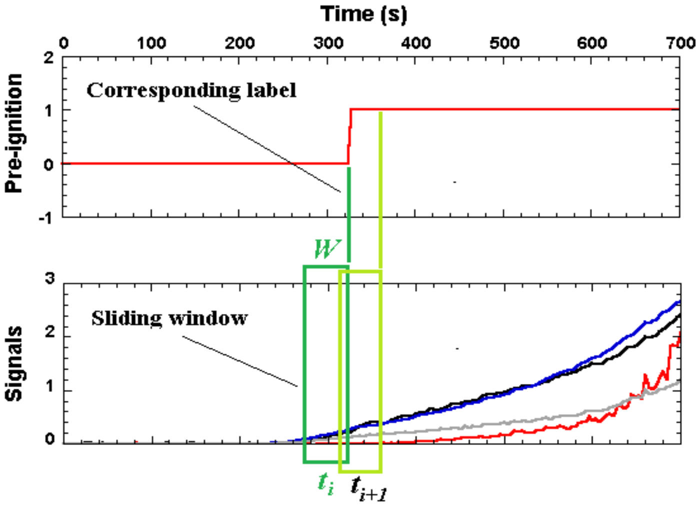


Figure 4. Schematic of moving windows with window size, W, and its corresponding label (the two sliding windows are not to scale; t_i and t_{i+1} are 4 s apart).

Table 1. Number of normal cooking and pre-ignition data instances for different groups of experiments when W = 60 s.

Group Label	Food type	Cooktop type	Normal cooking	Pre-ignition
OE (Oil, Electric)	Oil	Electric	2486	2249
OFE (Other Foods, Electric)	Other foods	Electric	2702	606
OG (Oil, Gas)	Oil	Gas	675	1022
All (OE+OFE+OG)	All	All	5863	3877

In addition to feature representation, each instance also needs to be well annotated. As shown in Fig 4, each data instance was labeled "*Normal cooking*" or "*Pre-Ignition*". For oils, the data is defined as "*Normal Cooking*" if the pan temperature is less than 300 ºC. Reference [13] provides detailed descriptions for the determination of normal condition for other foods, such as salmon, bacon, chicken, hamburgers, and french fries. Table 1 illustrates data distributions for different groups of experiments with W = 60 s. Depending on the size of time window (W), it should be noted that several instances (i.e. W x f – 1) will be lost.

3.3 Feature selections

After obtaining the processed data instances, features are extracted to build classifiers. Statistics, such as the mean, maximum, and standard deviation, are used to characterize a series of data or approximate the data distribution. These statistic-based representations may be useful to detect pre-ignition. For instance, pre-ignition is usually associated with high values of IAQ signal, so the maximum of the IAQ signal would be useful. We therefore propose a set of statistic-based features for unattended cooking detection.

Statistic-based features can represent the character of the data. However, they cannot capture the temporal or trend information. In order to extract this information, we also propose a set of trend-based features. Given the values of the signal within the data window, $S = [s_1, \dots, s_{n/2}, \dots, s_n]$, we extract trend-based features by computing: $s_n - s_1$, $s_{n/2} - s_1$, $s_n - s_{n/2}$, and max(S) – min(S). In addition to the raw signal, the first derivative is also calculated. We further extract the same set of features from the first derivative signal $S' = [s'_1, \dots, s'_{n/2}, \dots, s'_{n-1}]$, where $s'_i = s_{i+1} - s_i$. In total, 18 features for one signal can be extracted. As shown in Table 2, signals from 6 sensors, including alcohol, CO, dust, IAQ, smoke, and VOC are used. With these signals, a feature vector of 108 dimensions for each data instance is obtained. The feature vector is fed into a classifier to build the pre-ignition detection model.

Table 2. Signals and their features.

Sensors	Alcohol	CO	Dust	IAQ	Smoke	VOC
Signals	Raw signal (S) and the first derivative (S')					
Trend-based features	$s_n - s_1$, $s_{n/2} - s_1$, $s_n - s_{n/2}$, max(S) – min(S) $s'_{n-1} - s'_1$, $s'_{(n-1)/2} - s'_1$, $s'_{n-1} - s'_{(n-1)/2}$, max(S') – min(S')					
Statistical features	Mean, maximum, minimum, median, standard deviation					

3.4 Training and evaluation

The features introduced above are extracted for each instance, and then fed into a conventional machine learning algorithm to build the pre-ignition detection model. In this analysis, multiple machine learning algorithms are adopted for the detection model, including support vector machine (SVM), random forest (RF) [20], and decision tree (DT) [21]. These commonly used machine learning algorithms are reliable for many classification applications. Appendix C provides the basic concept and the model configuration for the three machine learning algorithms. Investigating and comparing the performance of these machine learning algorithms can help determine the most suitable machine learning model for ignition detection. In general, there is a tradeoff in real-time detection associated with the size of the moving window. Using a smaller moving window can increase the response time of the model, which is an important factor in real-time detection. On the other hand, using a larger moving window is likely to achieve higher accuracy, as the model has more information. Therefore, the impacts of algorithm type as well as window size are investigated here. Three window sizes are considered: 60 s, 32 s, and 16 s based on the frequency of the input signal (0.25 Hz).

For evaluation purposes, a leave-one-experiment-out cross-validation approach is used. The dataset is divided into N subsets, where N is the total number of experiments. Each subset contains all the data instances from one particular experiment, then the classifier is trained with N - 1 subsets and the classifier predictions are evaluated for the data instances in the remaining subset. This is repeated N times until all the subsets are evaluated once. Final performance is taken as the overall average accuracy across all data instances in the N evaluations.

4. Results and Discussion

4.1 Classification of unattended cooking for OE and OFE data

Data associated with all experiments[6] with oil on the electric cooktop (denoted as OE in Table 1) and other foods (OFE) on the electric cooktop were considered. Figure 5 shows an example of the prediction results of SVM with a moving window of 16 s for Exp 8 and Exp 46. The black curves are the classifier predictions. Blue curves are the converted values based on a discrimination threshold of 0.5. If the prediction value is less than 0.5, the prediction is classified as normal cooking. If the prediction is larger than 0.5, the prediction is classified as pre-ignition. As compared to the data label (red curves), the SVM tends to predict pre-ignition well before the ignition. In order to evaluate the performance over all experiments, the precision, recall, and F1-score measures are reported. Table 3 shows the prediction performance for the three moving window sizes. Here, precision is defined as the ratio of the number of true positives over the sum of true and false positives; recall is defined as the ratio of true positives over sum of true positives and false negatives. The F1 score is the weighted average of precision and recall.

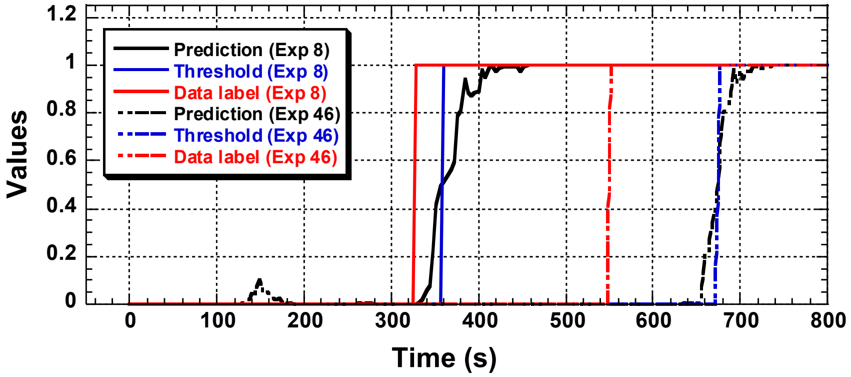


Figure 5. Performance comparison of SVM with W of 16 s for Exp 8 and Exp 46 considering both OE and OFE data.

[6] Exp 16, 39, 48, 49, 50, and 60 are excluded. Exp 39, 48, 50 and 60 are dual-pan experiments. Since single pan experiments are of interest, the dual-pan experiments will be considered in future study.

The Correct Classified Rates (CCR) for the three machine learning models with three moving window sizes are presented in Fig 6. CCR is defined as the ratio of the correct classified instances to the total instances in a dataset; it is an indicator of overall classification performance. The results in Fig. 6 suggest two general conclusions. First, the larger the window size, the better the observed performance. Second, SVM seems to have slightly higher CCRs compared to RF and DT. For SVM, prediction accuracies are 93.8 %, 92.8 %, and 91.2 % for *W* = 60 s, 32 s, and 16 s, respectively. This statistic demonstrates that with approximately 4 times the detection frequency, there was only a 2.5 % tradeoff on prediction accuracy.

Table 3. Precision, recall, and F1-score for SVM with three moving window sizes considering both OE and OFE data.

Window size	**Class**	**Precision**	**Recall**	**F1-score**
60 s	Normal	96.3%	94.1%	95.2%
	Unattended	89.7%	93.4%	91.5%
32 s	Normal	94.9%	93.8%	94.3%
	Unattended	89.3%	91.1%	90.1%
16 s	Normal	93.6%	92.4%	93.0%
	Unattended	87.2%	89.1%	88.1%

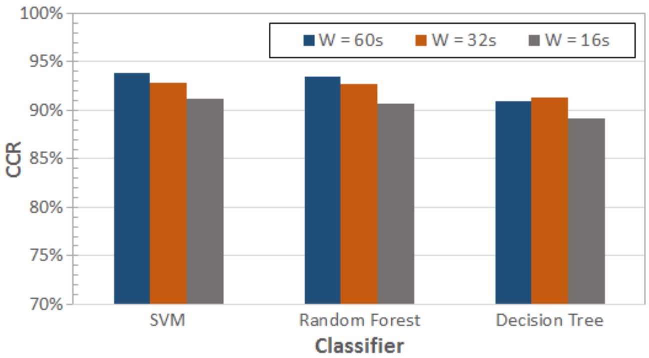


Figure 6. Overall performance classifying normal cooking and pre-ignition for three machine learning models with different moving window sizes considering both OE and OFE data.

4.2 Classification of pre-ignition using an object-specified approach

Sensors in the oil experiments, regardless of oil volume, oil type, and heating conditions, presented monotonic behavior. This trend does not necessarily exist for the cooking experiments with other foods due to a variety of factors including possibly the burner settings used, the water content of the foods, and the shape and deformation of the foods during cooking. If a machine learning model uses data that have similar behaviors, the prediction performance should increase. Therefore, an object-specified approach was also followed, training and evaluating the machine learning models either with only the OE data or only the OFE data.

Figure 7 shows the updated CCRs using the object-specified approach on SVM. For OE data alone, the prediction performance for *W* = 60 s is increased over the combined OE and OFE data to 96.9 %. In general, there is an average 1.9 % improvement in the prediction performance over the combined OE and OFE data for all three moving window sizes. For OFE data alone, the prediction performance for W = 16 s increases by 5.1 % over the combined OE and OFE data. The improved performance for OFE data alone with W = 60 s and W = 32 s is 2 % and 2.8 %, respectively. Table 4 shows the detailed breakdown of the two data sets using statistical measures, showing the enhanced prediction accuracy.

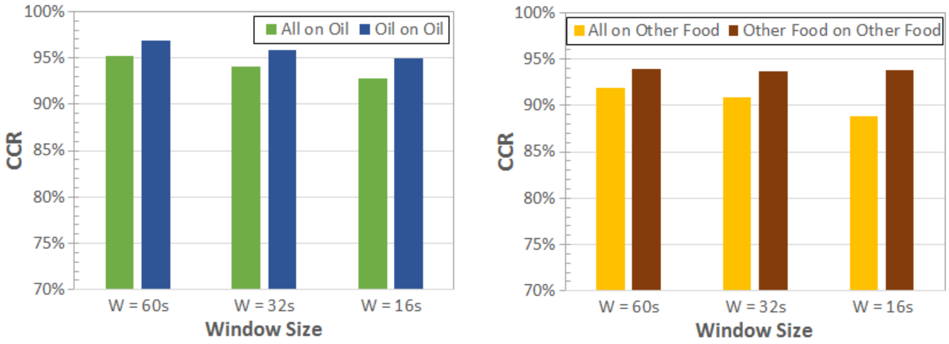


Figure 7. Performance of SVM predictions using both OE and OFE data for training compared to the object-specified approach, where only OE or OFE data is used for training.

Table 4. Precision, recall, and F1-score for SVM with three moving window sizes using the object-specified approach.

Dataset	Class	Precision	Recall	F1-score
Oil (OE)	Normal	98.4%	95.7%	97.0%
	Unattended	95.3%	98.2%	96.8%
Other Foods (OFE)	Normal	96.8%	95.8%	96.2%
	Unattended	82.0%	85.6%	83.8%

4.3. Towards Two-Step Pre-Ignition Detection

Based on the previous section, the object-specified machine learning models built to classify the cooking conditions (normal or pre-ignition) for a specific type of cooking (oil heated on an electric cooktop versus other foods on an electric cooktop) is shown to outperform the more generic machine learning model. This performance enhancement does not require additional data. Instead, to use the object-specified machine learning models, a classifier that discriminates between the oil cooking scenarios and other cooking scenarios is needed. The gas cooktop experiments were not included in the previous analyses, so cooktop type could also be important in determination an optimized model. Nevertheless, the feasibility of use of a two-step architecture for detection of pre-ignition is investigated.

Next, machine learning is used to detect the type of cooktop being used. The cooking oil results from different cooktop types in the dataset are used. The data from heating oil on an electric cooktop (OE) are given the label "*Electric*", and the data from heating oil on a gas cooktop (OG) are given the label "*Gas*". The same features as presented in Section 3.3 are used to build the cooktop type identification models and the same evaluation approach is adopted. As shown in Fig. 8, the attempt leads to very promising performance of cooktop identification. Using SVM as the classifier with W = 60 s, an accuracy of better than 98 % is observed. For a smaller W, which enables a faster time response, an accuracy of 98 % ($W = 32\ s$) and 97.3 % ($W = 16\ s$) is achieved. As shown in Table 5, the models achieve high F1-scores for both cooktop types, indicating that the model does not achieve high CCR by simply selection of the majority side. These results suggest that the proposed method can precisely differentiate an electric cooktop from a gas cooktop, providing verification of the machine learning methods developed here. The results also demonstrate that it may be possible to build a two-step pre-ignition detection model, which first classifies the cooking scenario to recommend a specific model to better predict pre-ignition conditions for a particular type of cooktop.

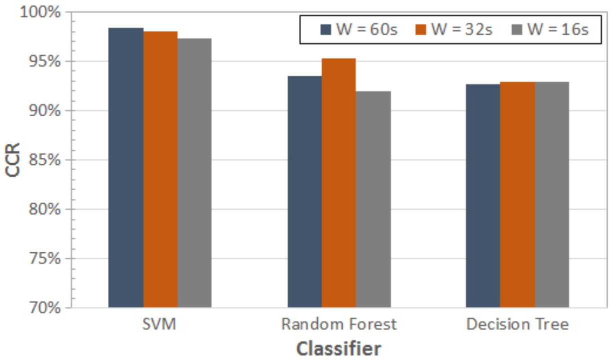


Figure 8. Overall performance on cooktop type classification for the three machine learning models with different *W* using OE and OG data sets.

Table 5. Precision, recall, and F1-score for SVM with three different *W* for cooktop type classification.

Window size	Class	Precision	Recall	F1-score
60 s	Electric	99.5%	98.4%	98.9%
	Gas	95.5%	98.5%	96.9%
32 s	Electric	98.9%	98.4%	98.7%
	Gas	95.4%	96.7%	96.0%
16 s	Electric	98.1%	98.2%	98.2%
	Gas	94.7%	94.5%	94.6%

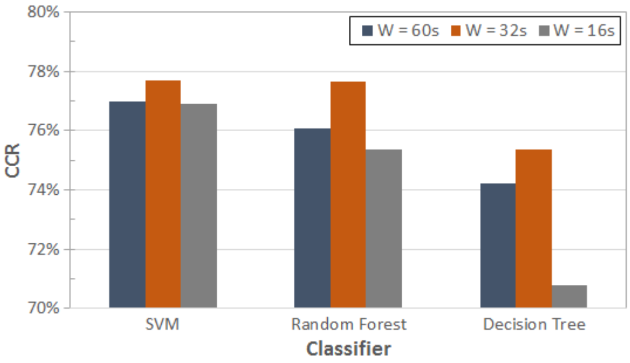


Figure 9. Overall performance on food type classification for the 3 machine learning models with different moving window size on OE and OFE data.

Tam, Wai Cheong; Fu, Eugene Yujun; Mensch, Amy; Hamins, Anthony; You, Christina; Ngai, Grace; Leong, Hong va. "Prevention of Cooktop Ignition Using Detection and Multi-Step Machine Learning Algorithms." Paper presented at International Association of Fire Safety Science 2020, Waterloo, CA. April 27, 2020 - May 01, 2020.

Next, the possibility of using the classifier to identify the heating of oil versus cooking scenarios involving foods is investigated. Both the OE and OFE data sets using the electric cooktop are utilized. The data with oil (OE) is labeled "*Oil*" and the data with other foods (OFE) is labeled "*Others*". The same model features and evaluation approach presented in Section 3.3 to build the oil versus other scenario detection models are adopted here. Table 6 provides the detailed statistical performance for the food type predictions. The numerical results indicate that the models cannot precisely differentiate cooking oil from scenarios involving cooking other foods. A possible reason is that there is insufficient data. The models to detect cooktop type achieved promising performance despite fewer data sets using the gas cooktop because the sensor signals from heating oil with a gas cooktop does not vary as much across different experiments as the varied data sets in the OFE data group. The types of other foods and procedures of cooking other foods are much more diverse than just heating oil on the highest setting, leading to diverse sensor signal patterns that make it difficult for the classifier to learn patterns.

Table 6. Precision, recall, and F1-score for SVM with three moving window sizes on classifying heating oil from other cooking scenarios on electric cooktops.

Window size	Class	Precision	Recall	F1-score
60 s	Oil	74.9%	92.6%	82.8%
	Others	82.9%	53.4%	65.0%
32 s	Oil	75.0%	94.5%	83.6%
	Others	86.2%	51.9%	64.8%
16 s	Oil	74.2%	94.6%	83.2%
	Others	85.7%	49.6%	62.8%

5. Conclusions

The feasibility of building machine learning models to perform real-time cooktop pre-ignition detection is investigated. Machine learning algorithms have the capability to consider multiple sensor signals. Taking advantage of that capability, statistic-based and trend-based features are extracted from the time series signal of six sensors: alcohol, CO, dust, IAQ, smoke, and VOC, to build pre-ignition detection models. The proposed approach achieves encouraging performance, even when using data from diverse cooking scenarios on electric cooktops (OE and OFE data sets) for training with 93.8 % of data instances predicted correctly (using SVM and $W = 60$ s).

Models trained and tested only on data for a specified cooking condition (separating OE and OFE data sets) outperform models trained on the combined set of all electric cooktop data. For instance, the overall accuracy is 96.9 % for the model trained and tested on only the OE data set (using SVM and $W = 60$ s). If the cooking scenario of the target data can be identified, the detection performance of pre-ignition can be improved. This suggests the potential of a two-step approach to obtain a more robust cooking pre-ignition detection model.

Results from calculations to identify cooking scenarios using a multi-step detection approach shows it is possible to precisely differentiate heating oil on electric cooktop from a gas cooktop. However, the method was not as effective in identifying a scenario of heating oil versus cooking other foods on an electric cooktop. In the future, experiments on a wider variety of food types will be considered to test the range of possible improvement of model performance.

Acknowledgement

The authors are very grateful to Dr. Michael Huang for helpful discussions and suggestions.

Supplementary material

Sensor data associated with this article and that of obtained from [12,13] can be found at: https://doi.org/10.18434/M32171.

References

[1] Ahrens, M., 2009. *Home fires involving cooking equipment*. Quincy: National Fire Protection Association.

[2] Underwriter's Laboratory, 2014 and 2018. Standard for Household Electric Ranges (UL 858). Northbrook IL.

[3] Hamins, A., Kim, S.C. and Madrzykowski, D., 2018. Characterization of stovetop cooking oil fires. *J Fire Sciences*, *36*(3), pp.224-239.

[4] Hamins, A., Kim, S.C. and Bundy, M., 2018. *Investigation of Residential Cooktop Ignition Prevention Technologies*. US Department of Commerce, National Institute of Standards and Technology.

[5] Gottuk, D.T., Wright, M.T., Wong, J.T., Pham, H.V. and Rose-Pehrson, S.L., 2002. *Prototype Early Warning Fire Detection System: Test Series 4 Results* (No. NRL/MR/6180--02-8602). NAVAL RESEARCH LAB WASHINGTON DC.

[6] Cestari, L.A., Worrell, C. and Milke, J.A., 2005. Advanced fire detection algorithms using data from the home smoke detector project. *Fire Safety Journal*, *40*(1), pp.1-28.

[7] Jain, A., Nyati, P., Nuwal, N., Ansari, A., Ghoroi, C.H.I.N.M.A.Y. and Ghandi, P., 2014. Pre-Detection of Kitchen Fires due to Auto-Ignition of Cooking Oil and LPG Leakage in Indian Kitchens. *Fire Safety Science*, *11*, pp.1285-1297.

[8] Johnsson, E. and Zarzecki, M., 2017. Using Smoke Obscuration to Warn of Pre-Ignition Conditions of Unattended Cooking Fires. In *16th International Conference on Automatic Fire Detection (AUBE'17) & Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2017).*

[9] Gaur, A., Singh, A., Kumar, A., Kulkarni, K.S., Lala, S., Kapoor, K., Srivastava, V., Kumar, A. and Mukhopadhyay, S.C., 2019. Fire Sensing Technologies: A Review. *IEEE Sensors Journal*, 19(9), pp.3191-3202.

[10] Fonollosa, J., Solórzano, A. and Marco, S., 2018. Chemical sensor systems and associated algorithms for fire detection: A review. *Sensors*, 18(2), p.553.

[11] Johnsson, E.L., 1998. Study of technology for detecting pre-ignition conditions of cooking-related fires associated with electric and gas ranges and cooktops, final report, NISTIR 5950, National Institute of Standards and Technology, Gaithersburg, Maryland.

[12] Mensch, A., Hamins, A. and Markell, K., 2018, September. Development of a Detection Algorithm for Kitchen Cooktop Ignition Prevention. In *Suppression, Detection and Signaling Research and Applications Conference Proceedings (SUPDET 2018)*.

[13] Mensch, A., Hamins, A., Lu, J., Kupferschmid, M., Tam, W.C. and You, C., 2019, September. Evaluating Sensor Algorithms to Prevent Kitchen Cooktop Ignition and Ignore Normal Cooking. In *Suppression, Detection and Signaling Research and Applications Conference Proceedings (SUPDET 2019)*.

[14] Jang, D.W., Lee, S., Park, J.W. and Baek, D.C., 2018. Failure detection technique under random fatigue loading by machine learning and dual sensing on symmetric structure. *International Journal of Fatigue*, 114, pp.57-64.

[15] Eskandarpour, R., Khodaei, A. and Arab, A., 2017, September. Improving power grid resilience through predictive outage estimation. In *2017 North American Power Symposium (NAPS)* (pp. 1-5). IEEE.

[16] Bhattacharya, B. and Sinha, A., 2017, November. Intelligent fault analysis in electrical power grids. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 985-990). IEEE.

[17] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., 1998. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4), pp.18-28.

[18] Popov, P.P., Buchta, D.A., Anderson, M.J., Massa, L., Capecelatro, J., Bodony, D.J. and Freund, J.B., 2019. Machine learning-assisted early ignition prediction in a complex flow. *Combustion and Flame*, 206, pp.451-466.

[19] Mensch, A., Hamins, A., Tam, W.C., Lu, J., Markell, K., You, C., and Kupferschmid, M., Sensors and Machine Learning Models to Prevent Cooktop Ignition and Ignore Normal Cooking. *Fire Technology (Submitted).*

[20] Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

[21] Kwok, S.W. and Carter, C., 1990. Multiple decision trees. In *Machine Intelligence and Pattern Recognition* (Vol. 9, pp. 327-335). North-Holland.

[22] Mishra, M., & Rout, P. K., 2017. Detection and classification of micro-grid faults based on HHT and machine learning techniques. *IET Generation, Transmission & Distribution*, *12*(2), 388-397.

[23] Kazem, H. A., Yousif, J. H., & Chaichan, M. T., 2016. Modeling of daily solar energy system prediction using support vector machine for Oman. *International Journal of Applied Engineering Research*, *11*(20), 10166-10172.

[24] Moutis, P., Skarvelis-Kazakos, S., & Brucoli, M., 2016. Decision tree aided planning and energy balancing of planned community microgrids. *Applied Energy*, *161*, 197-205.

[25] Jiang, H., Li, Y., Zhang, Y., Zhang, J. J., Gao, D. W., Muljadi, E., & Gu, Y., 2017. Big data-based approach to detect, locate, and enhance the stability of an unplanned microgrid islanding. *Journal of Energy Engineering*, *143*(5), 04017045.

[26] Vapnik, V., 1998. The support vector method of function estimation. In *Nonlinear Modeling* (pp. 55-85). Springer, Boston, MA.

[27] Breiman, L., 2017. *Classification and regression trees*. Routledge.

[28] Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

[29] Oliphant, T.E., 2007. Python for scientific computing. *Computing in Science & Engineering*, *9*(3), pp.10-20.

[30] Staelin, C., 2003. Parameter selection for support vector machines. *Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1*, *1*.

[31] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), pp.2825-2830.

Appendix A. Additional Information for Sensor Array.

14 different sensor responses were selected for testing. The sensors were based on various operating mechanisms, including electrochemical, MOS-type, light scattering, and non-dispersive infrared absorption. Sensors were selected to measure CO_2, CO, hydrocarbons, alcohols, H_2, natural gas, volatile organic compounds (VOCs), smoke, air quality, and aerosols/dust. Humidity and temperature were also measured. Table A1 provides the sensor names, their sensitivity, operating principle, measurement range, and units.

Table A1. Summary of sensor information.

	Sensor Name	Sensitivity	Operating Principle	Measurement Range	Units
1	Smoke	combustible gas, smoke	electrochemical	300 -10,000	ppm
2	Alcohol	alcohol	electrochemical	0.04 - 4	mg/L
3	Hydrocarbon 1	methane, propane, butane	electrochemical	300 -10,000	ppm
4	Hydrocarbon 2	liquified petroleum, butane, propane, LPG	electrochemical	300 - 10,000	ppm
5	Hydrogen	H_2	electrochemical	100 - 1,000	ppm
6	Natural Gas	methane	electrochemical	< 10,000	ppm
7	CO 1	CO	electrochemical	0 - 10,000	ppm
8	VOCs	air contaminants, VOCs, odorous gases	metal oxide sensor	not specified	ppm
9	Dust	aerosol	optical	0.1 - 0.5	mg/m^3
10	Humidity	H_2O	electrochemical	2000	ppm
11	CO_2 1	CO_2	electrochemical	2000	ppm
12	IAQ	cooking odors, pollutants, smoke	electrochemical	not specified	ppm
13	CO_2 2	CO_2	electrochemical	5000	ppm
14	CO 2	CO	electrochemical	5000	ppm

Appendix B. Summary of important experimental information.

EXP	Ignition	Pan Type and Size (cm)	Food Type and Amount	Burner Type and Size	Hood Flow	Foil Surround
1	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
2	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
3	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
4	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
5	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
6	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
7	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
8	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
9	Yes	Cast Iron, 20	Canola Oil, 100mL	Electric, Small	High	No
10	Yes	Aluminum, 20	Canola Oil, 50mL	Electric, Small	High	Yes
11	Yes	Multi-layered, 20	Canola Oil, 50mL	Electric, Small	High	Yes
12	Yes	Stainless Steel, 20	Canola Oil, 50mL	Electric, Small	High	Yes
13	Yes	Cast Iron, 20	Canola Oil, 200mL	Electric, Small	High	Yes
14	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Big	High	Yes
15	Yes	Cast Iron, 25	Canola Oil, 100mL	Electric, Big	High	Yes
16	No	Aluminum, 20	Corn Oil, 50mL	Electric, Small	High	No
17	Yes	Aluminum, 20	Corn Oil, 50mL	Electric, Small	High	No
18	Yes	Cast Iron, 20	Corn Oil, 50mL	Electric, Small	High	No
19	Yes	Cast Iron, 25	Corn Oil, 100mL	Electric, Big	High	Yes
20	Yes	Cast Iron, 20	Corn Oil, 50mL	Electric, Small	High	Yes
21	Yes	Cast Iron, 20	Soy Oil, 50mL	Electric, Small	High	Yes
22	Yes	Cast Iron, 25	Soy Oil, 100mL	Electric, Big	High	Yes
23	Yes	Cast Iron, 20	Olive Oil, 50mL	Electric, Small	High	Yes
24	Yes	Cast Iron, 25	Olive Oil, 100mL	Electric, Big	High	Yes
25	Yes	Cast Iron, 25	Sunflower Oil, 100mL	Electric, Big	High	Yes
26	Yes	Cast Iron, 20	Sunflower Oil, 50mL	Electric, Small	High	Yes
27	Yes	Cast Iron, 20	Butter, 45.68g	Electric, Small	High	Yes
28	No	Broiler Pan	Hamburger, 1.14kg	Oven	High	NA
30	No	Cast Iron, 25	Hamburger, 1.14kg	Electric, Big	High	Yes
31	Yes	Cast Iron, 20	Salmon, 18oz & Butter 42.5g	Electric, Small	High	Yes
32	No	Cast Iron, 25	Salmon, 2.8oz & Butter 85.1g	Electric, Big	High	Yes
33	No	Cast Iron, 20	Water, 50mL	Electric, Small	High	Yes
34	No	NA	NA	Electric, Big	High	NA
35	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	High	No
36	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	Medium	No
37	No	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	Medium	No
38	No	Aluminum, 20	Canola Oil, 50mL	Electric, Small	Medium	No
39	No	Cast Iron, 20 & Aluminum, 20	Canola Oil, 50mL & Canola Oil, 50mL	Electric, Big & Electric, Small	Medium	No
40	No	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	Off	No
41	No	Broiler Pan	Hamburger, 1.14kg	Oven	High	NA

EXP	Ignition	Pan Type and Size (cm)	Food Type and Amount	Burner Type and Size	Hood Flow	Foil Surround
42	No	Cast Iron, 20	Salmon, 18oz & Butter 42.4g	Electric, Small	High	No
43	No	Cast Iron, 20	Chicken legs, 2 pieces & Canola Oil, 200mL	Electric, Small	High	No
44	Yes	Cast Iron, 25	French fries, 223.3g & Canola Oil, 500mL	Electric, Big	High	No
45	No	Cast Iron, 25	Bacon, 228g	Electric, Big	Medium	No
46	Yes	Cast Iron, 20	Bacon, 110g	Electric, Small	Medium	No
47	No	Cast Iron, 25	Hamburger, 1.14kg	Electric, Big	Medium	No
48	Yes	Cast Iron, 20 & Cast Iron, 25	Canola Oil, 50mL & Canola Oil, 100mL	Electric, Big & Electric, Small	High	No
49	Yes	Cast Iron, 20 & Cast Iron, 25	Canola Oil, 50mL & Canola Oil, 100mL	Electric, Big & Electric, Small	High	No
50	Yes	Cast Iron, 20 & Cast Iron, 25	Canola Oil, 50mL & Canola Oil, 100mL	Electric, Big & Electric, Small	High	No
51	Yes	Cast Iron, 20	Canola Oil, 50mL	Electric, Small	Low	No
52	No	NA	NA	Gas, Big	Medium	No
53	No	Cast Iron, 20	Canola Oil, 50mL	Gas, Medium	Medium	No
54	Yes	Cast Iron, 25	Canola Oil, 100mL	Gas, Big	Medium	No
55	No	Cast Iron, 25	NA	Gas, Big	Medium	No
56	Yes	Cast Iron, 25	Canola Oil, 100mL	Gas, Big	Medium	No
57	Yes	Cast Iron, 20	Canola Oil, 50mL	Gas, Big	Medium	No
58	Yes	Cast Iron, 25	Canola Oil, 100mL	Gas, Big	Medium	No
59	No	NA	NA	Gas, Big	Medium	No
60	No	Cast Iron, 20 & Cast Iron, 25	Canola Oil, 50mL & Canola Oil, 100mL	Gas, Big & Gas, Medium	Medium	No

It should be noted that all sensor data listed in this appendix can be found at https://doi.org/10.18434/M32171.

Appendix C. Basic Concept for Support Vector Machine, Decision Tree, and Random Forest.

Machine learning (ML) algorithms have been widely used for multi-class classification problems in various fields. Based on recent literature [22-25], it has been demonstrated that support vector machine (SVM), decision tree (DT), and random forest (RT) have the capabilities to handle complex time series data with multi-dimensional feature vectors. Guidelines are lacking for the use of ML algorithms in classification problems involving time series data in fire research, and the performance for these three ML algorithms is unknown. Therefore, SVM, DT, and RF will be used for the development of classification models for prediction of fire hazards. Comparing the results obtained from the different models can serve as a sanity check for the performance of each model. In the next subsection, the basic concepts for SVM, DT, and RT will be presented. Readers can refer to the following references [26-28] for detailed descriptions of the mathematical formulation for each algorithm.

<u>Support Vector Machine (SVM) [26]</u>

SVM is a classifier that finds a decision boundary, known as a hyperplane, to separate instances into two classes. The algorithm maximizes the constrained margin such that the distance between the instances in different classes is optimized to achieve the greatest model generalizability. For example, given a training dataset $T = \{(X_1, y_1), (X_2, y_2), \dots, (X_n, y_n)\}$, which can be linearly separated, the hyperplane denoted as p can be written as:

$$w \bullet X + b = 0 \tag{C1}$$

where X_n is the sample of n^{th} instance, and y_n is the class label. w is the weight of the hyperplane, and b is the bias of the hyperplane. Based on the definition provided in [26], the distance between the instances for different classes is:

$$d = \min_{i=1,2,\dots,n} y_i(\frac{w}{\|w\|} \cdot X_i + \frac{b}{\|w\|}) \tag{C2}$$

where $\|w\|$ is norm of w. For SVM, the distance is known as margin. Therefore, SVM determines the hyperplane with the largest margin by solving the optimization problem:

$$\arg\max_{w,b} (\min_{i=1,2,\dots,n} y_i(\frac{w}{\|w\|} \cdot X_i + \frac{b}{\|w\|})) \tag{C3}$$

For real-life applications, fire data are often more complex and not linearly separable. In order to overcome this numerical difficulty, there are two treatments. The first treatment is called the “kernel trick” [26], which commonly involves four nonlinear kernel functions: 1) polynomial kernel, 2) Gaussian kernel, 3) radial basis function, and 4) sigmoid kernel. The use of a kernel function allows the transformation of data into a higher dimensional space such that a hyperplane exists separating the instances X_n for different classes. The second treatment is introducing a regularization or slack variable. With the implementation of the regularization variable, a small proportion of the data are ignored, and misclassification is allowed. Although there is trade-off for use of the regularization variable treatment, it generally helps to avoid over-fitting and to provide a more generalized model.

Decision Tree (DT) [27]

DT builds classification models in the form of a tree structure. A typical DT is composed of a root node, internal nodes, edges, and leaves. The root node represents the entire population. Internal nodes represent partitioning or splitting conditions, which correspond to a feature vector. Edges can be a specific value or range of values for the splitting condition of the feature. The leaves represent the terminal nodes of a tree with class labels. The hierarchical nature of the algorithm provides detailed information of how a decision is being made. As compared to SVM, DT is more transparent, and the results are easier to interpret.

Given the aforementioned training dataset $T = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, the formulation of a DT involves selecting optimal splitting features. The process starts by splitting the dependent feature, or root node, into binary pieces, where the child nodes have less entropy than the parent node. If the sample is completely homogeneous the entropy is zero, and if the sample is equally divided the entropy is one. Mathematically, entropy is defined as:

$$Entropy = -\sum_{i=i}^{k} P_k log P_k \tag{C4}$$

where P_k is possibility of the instance belonging to class k. In general, DT searches through all candidate splits to find the optimal split that minimizes the resulting entropy of a tree. One effective split strategy is to always select the feature with largest information gain, IG:

$$IG(D_p, f) = Entropy(D_p) - \sum_{i}^{all\ child\ node} \frac{N_i}{N} Entropy(D_i) \tag{C5}$$

where D_p is the dataset of the parent node with N number of samples, f is the feature, and D_i is the dataset of the i^{th} child node with N_i number of instances. This splitting process is continued until all the instances at current level are labeled to the appropriate classes.

Random Forest (RF) [28]

RF is an ensemble learning method for classification. It builds H number of classification trees and provides the prediction of the class of an object based on the averaged results obtained from each of the trees. Mathematically, after H trees are grown, the RF classification predictor is given as:

$$f(x) = \frac{1}{H} \sum_{i=1}^{H} K(i, x) \tag{C6}$$

where x is the input feature.

In general, significant effort is usually needed to tune the model to maximize performance. This process can be accomplished by selecting appropriate hyperparameters, which can be thought of as the “dials” or “knobs” of a machine learning model. There are several automated tuning methods, such as grid search, random search, and Bayesian optimization. In this study, the most basic tuning method, grid search [29], is used. With this technique, we simply build a model for each possible combination of all the hyperparameter values provided, evaluate each model, and select the architecture which produces the best results.

Summary

Table C1 provides the summary of model configurations for the three ML algorithms. For SVM, the radial basis functions (RBF) kernel function is used. The selection of the kernel function depends on both the number of input features and the size of the dataset. In this study, there are up to 12 features and roughly 10000 instances (refer to Table 1). Since the number of samples from the dataset is much larger than the number of input features, it is suggested that RBF will have better performance [30]. With the use of the RBF kernel, the two parameters, C and γ, must be considered. The parameter C is the regularization parameter or slack variable, and it controls the trade-off between misclassification of training examples and simplicity of the decision surface. In general, a low C makes the decision surface smooth, while a high C aims to classify all training examples correctly [31]. The parameter γ defines how much influence a single training example has on other examples. The larger γ is, the less influence a single simple will have. Grid search [29] is used to determine the optimal values for the two parameters.

Table C1: Summary of model configurations for SVM, DT, and RF.

SVM				DT	RF			
	Optimal	**Range**	**Interval**			**Optimal**	**Range**	**Interval**
C	100	0 - 300	10	NA	**Estimator**	10	5 - 100	5
Gamma (γ)	10	0 - 50	1					

The model configurations for DT and RF are simpler. For DT, the entropy method is used to measure the quality of a split, but no parameters were needed to be adjusted. For RF, the setting remains similar except that the optimal number of trees needs to be identified. In theory, the more trees the better. However, based on results obtained for different number of trees ranging from 10 to 100 with an interval of 10 trees, the improvement of prediction accuracy with more trees is negligible (less than 0.1 %). Therefore, the number of trees is set to 10.

Figure captions

Figure 1. Schematic drawing of experimental setup (not to scale) and detailed view of the sensor array (in the duct).

Figure 2. Signals for 11 selected sensors and pre-ignition condition for Exp 8.

Figure 3. Comparison of IAQ and VOCs signals for 3 different tests (Exp 8, 46, and 57).

Figure 4. Schematic of moving windows with window size, *W*, and its corresponding label (the two sliding windows are not to scale; t_i and t_{i+1} are 4 s apart).

Figure 5. Performance comparison for Exp 8 and Exp 46 on OE and OFE data.

Figure 6. Overall performance on normal/unattended cooking classification for 3 machine learning models with different moving window size on OE and OFE data.

Figures 7. Performance improvement using the proposed object-oriented approach.

Figure 8. Overall performance on stove type classification for the 3 machine learning models with different moving window size on OE and OG data.

Figure 9. Overall performance on food type classification for the 3 machine learning models with different moving window size on OE and OFE data.

Streaming Batch Gradient Tracking for Neural Network Training (Student Abstract)

Siyuan Huang,[1] **Brian D. Hoskins,**[2] **Matthew W. Daniels,**[2] **Mark D. Stiles,**[2] **Gina C. Adam**[1]*
[1]School of Engineering & Applied Science, George Washington University, Washington, DC, USA
[2]Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

Abstract

Faster and more energy efficient hardware accelerators are critical for machine learning on very large datasets. The energy cost of performing vector-matrix multiplication and repeatedly moving neural network models in and out of memory motivates a search for alternative hardware and algorithms. We propose to use streaming batch principal component analysis (SBPCA) to compress batch data during training by using a rank-k approximation of the total batch update. This approach yields comparable training performance to minibatch gradient descent (MBGD) at the same batch size while reducing overall memory and compute requirements.

Introduction

Recent assessment of the energy necessary to train deep networks highlights the high financial and environmental costs associated with this fast growing field (Strubell, Ganesh, and McCallum 2019). New hardware accelerators are needed for efficient local and cloud computing (Ambrogio et al. 2018).

To accelerate training, research has increasingly focused on variable learning rate methods that add additional hyperparameters to the training of neural networks. Though these additional hyperarameters decrease the training time, they double or triple the memory overhead compared to a unit batch size stochastic gradient descent (SGD). These approaches are nevertheless preferred over SGD, however, since they can accelerate training on the rapidly fluctuating gradient vector to provide a superior learning trajectory.

In our approach, we accelerate training and reduce overhead by generating a stochastic low-rank approximation of the gradient using streaming principal component analysis (SPCA). SPCA was proposed by Oja in 1982(Oja 1982), and recent proposals combine SPCA with adaptive algorithms to optimize the learning rate while using only a single pass over the data (Henriksen and Ward 2019). (Burrello et al. 2019) proposed a parallel implementation of streaming History-PCA to save memory when backpropagating. SPCA remains an active research area which is promising for edge computing and other applications where memory usage is critical. To our knowledge, Hoskins et al. are first to use SPCA to compress batch gradients (Hoskins et al. 2019). Here we propose the SBPCA training algorithm, which approximates minibatch gradient descent with a slowly varying thin singular value decomposition of the overall gradient. Initial results show that SBPCA can accelerate training while reducing memory overhead and energy costs.

Method Details

In any SGD-derived method, the weights in a given layer are updated as $\Theta \leftarrow \Theta - \alpha\nabla\Theta$ where Θ is the weight matrix trained for that layer, α is the learning rate, and $\nabla\Theta$ is a stochastic approximation of $\partial\ell/\partial\Theta$ over a batch B, with ℓ the loss function evaluated after the feedforward step.

We consider an alternative calculation of $\nabla\Theta$ through a streaming low rank approximation based on the Singular Value Decomposition (SVD). Using Algorithm 1, the stochastic approximation is given by $\nabla\Theta = X\Sigma\Delta^T$, where X is the left singular matrix, σ is the vector of singular values, $\Sigma = \text{diag}(\sigma)$, and Δ is the right singular matrix. Restricting σ to its top k singular values, the memory cost of Algorithm 1 is $m \times k + k + n \times k$. We generally take $k \ll \min\{m, n\}$ to approximate $\nabla\Theta$ with efficient memory usage. Algorithm 1 updates this approximation using a novel block-averaged bi-iterative implementation of Oja's rule, and uses QR decomposition to re-orthogonalize the singular vectors.

```
Algorithm 1 Streaming Batch PCA (SBPCA) Update
Require: $\sigma$, $X$, $\Delta$ and $b$
  for $i = 1, 2, ..., B/b$ do
    $X$ step:                                   $\Delta$ step:
    $y = \delta_i\Delta/b$                       $z = x_i X/b$
    $X \leftarrow \frac{iX}{i+1} + \frac{x_i^T y}{(i+1)\sigma}$       $\Delta \leftarrow \frac{i\Delta}{i+1} + \frac{\delta_i^T z}{(i+1)\sigma}$
    $X \leftarrow \text{QR}(X)$                  $\Delta \leftarrow \text{QR}(\Delta)$
    $X$-$\Delta$ step: $\sigma \leftarrow \frac{i\sigma}{i+1} + \sum_{\text{rows}} \frac{(\delta_i\Delta)\odot(x_i X)}{(i+1)b}$
  end for
  Calculate $\nabla\Theta = X \cdot \text{diag}(\sigma) \cdot \Delta^T$
```

The X matrix (and Δ respectively) is updated as $X \leftarrow cX + (1-c)d$ where c is a convergence coefficient and d represents the block-averaged update of X over a minibatch of the input data, represented as x (and δ respectively), a matrix

[*] Correspondence: GinaAdam@gwu.edu

Huang, Siyuan; Hoskins, Brian; Daniels, Matthew; Stiles, Mark; Adam, Gina. "Streaming Batch Gradient Tracking for Neural Network Training." Paper presented at AAAI Conference on Artificial Intelligence, New York, NY, US. February 07, 2020 - February 12, 2020.

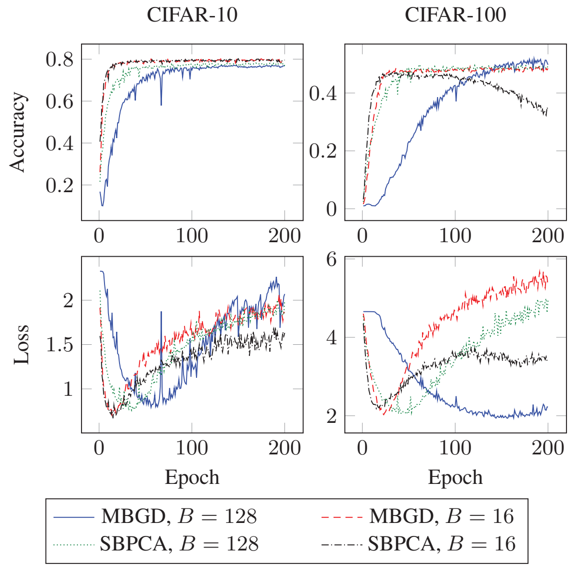


Figure 1: Comparative accuracy and loss for SBPCA and MBGD for CIFAR-10 (rank $k = 3$) and CIFAR-100 ($k = 30$) for $B \in \{16, 128\}$, with $b = B/4$. SBPCA recreates the loss minimum artifact in AlexNet present in MBGD.

with with rows b representing a subset of a minibatch of data with block size b examples. We describe two approaches. In the fixed version of Algorithm 1, we vary the convergence coefficient from block to block as $c_i = i/(i+1)$ while keeping the block size b of the minibatch fixed. However, we find that fixed values of b can lead to poor sampling of the space, so we introduce an alternative version (SBPCA with variable b or SBPCAV), adapted from Algorithm 1, with fixed $c = 1/2$ (implementable in hardware as a single bit-shift) but variable block size $b_i = 2^{(i-1)}$, with i the block index. The rank-1 updates corresponding to each singular vector of X and Δ are rescaled by their respective values in σ.

Experiments

We evaluate the SBPCA and SBPCAV algorithms on the CIFAR-10, CIFAR-100, and ImageNet datasets using an AlexNet modified to accommodate the smaller input size while maintaining the five convolutional and three fully connected layer structure. These algorithms are used only for fully connected layers; convolutional layers have less memory overhead, and so benefit less. We compare with MBGD results.

Since the initialization condition of the gradient approximation for the next batch is the end condition of the prior batch, the gradient estimation includes significant prior gradient history. This acts as a form of momentum, accelerating the training convergence compared to MBGD. In our experiments, we observed this is especially powerful at low learning rates. At high learning rates, the gradient changes more rapidly than the gradient estimation can be updated, destabilizing the training process. As seen in Fig. 1, the SPCA consistently achieves faster convergence than the classic MBGD for equivalent learning rates. Since AlexNet's structure poorly solves CIFAR, it leads to a non-monotonic behavior in the loss due to overtraining and can converge to an undesired local minimum. For small datasets, SBPCA approaches yield-equivalent accuracy to MBGD, shown in Table 1. Even for ImageNet, the algorithm can reach 85% of MBGD accuracy. We also see a dramatic difference in the performance of SBPCAV compared to SBPCA.

Table 1: Accuracy of training methods and ranks for CIFAR-10, CIFAR-100, and ImageNet. CIFAR hyperparameters same as before. For ImageNet, $B = 256$ and $b = 64$. *Fixed* refers to SBPCA and *Varied* to SBPCAV.

	CIFAR-10		**CIFAR-100**		**ImageNet**	
Rank	*Fixed*	*Varied*	*Fixed*	*Varied*	*Fixed*	*Varied*
1	0.7644	0.7391	0.4103	0.4163	0.3382	0.3660
3	0.7817	0.7729	0.4288	0.3783	0.3747	0.4065
10	0.7913	0.7840	0.4652	0.4252	0.4329	0.4454
30			0.4988	0.4563	0.4698	0.4602
100			0.5159	0.4712		
MBGD:	**0.7712**		**0.5185**		**0.5434**	

Conclusions

Our proposed SBPCA and SBPCAV can produce stochastic approximations of the gradient updates sufficient to train functionally relevant neural networks. We find that approximations of rank less than about ten are typically good enough to capture most relevant information about the local loss function. We verify these methods' effectiveness on three image datasets. While results are better on CIFAR-10 and CIFAR-100, even on ImageNet, SBPCAV can reach 85% of MBGD accuracy at significantly lower memory overhead. These results suggest future research exploring detailed memory analysis, impact of dropout and low rank approximations of more sophisticated training algorithms.

References

Ambrogio, S.; Narayanan, P.; Tsai, H.; et al. 2018. Equivalent-accuracy accelerated neural-network training using analogue memory. *Nature* 558(7708):60–67.

Burrello, A.; Marchioni, A.; Brunelli, D.; et al. 2019. Embedding principal component analysis for data reduction in structural health monitoring on low-cost iot gateways. *International Conference on Computing Frontiers* 235–239.

Henriksen, A., and Ward, R. 2019. Adaoja: Adaptive learning rates for streaming pca. *arXiv:1905.12115*.

Hoskins, B. D.; Daniels, M. W.; Huang, S.; et al. 2019. Streaming batch eigenupdates for hardware neural networks. *Frontiers in neuroscience* 13:793.

Oja, E. 1982. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology* 15(3):267–273.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Huang, Siyuan; Hoskins, Brian; Daniels, Matthew; Stiles, Mark; Adam, Gina. "Streaming Batch Gradient Tracking for Neural Network Training." Paper presented at AAAI Conference on Artificial Intelligence, New York, NY, US. February 07, 2020 - February 12, 2020.

Aggregating Atomic Clocks for Time-Stamping

T. Saidkhodjaev (UMD), J. Voas (NIST), R. Kuhn (NIST), J. DeFranco (PSU), and P. Laplante (PSU)

Abstract

A timestamp is a critical component in many applications, such as proof of transaction ordering or analyzing algorithm performance. This paper reports on a method called Verified Timestamping (VT) that improves the standard timestamp protocol. VT was developed at the National Institute of Standards and Technology (NIST) for use in algorithms where timestamp accuracy is critical. VT is an aggregation of the outputs from various atomic clocks to create a Timestamping Authority (TsA). The motivation for this research effort included malicious delay issues in Networks of Things [NIST SP 800-183] as well as race conditions associated with the inclusion of new blocks into blockchains. This paper presents the TsA design and the results of VT, which indicate that atomic clock aggregation is not only possible, but a viable means to produce higher integrity timestamps at the ms level of performance. Tests showed that this is sufficient to preserve event ordering, using only a conventional PC with no dedicated connection or specialized hardware.

1.0 Introduction

The term “timestamping” refers to marking the time when a certain event occurred, such as when a message was sent or received. Unfortunately, computer clocks are generally not precise enough for some types of data, such as in a financial transaction where it is essential to determine if a stock purchase occurred during the period of a particular price, for example. In the current environment of high frequency trading, accuracy on the order of µs is needed for audit and market surveillance by regulators (SR FINRA 2016-005). European Commission regulations announced in 2016 for the accuracy of business clocks require clock granularity of one µs or better, with a maximum allowed divergence from Coordinated Universal Time (UTC)

Saidkhodjaev, Temur; Voas, Jeff; Kuhn, D. Richard; DeFranco, Joanna; Laplante, Phil. ”Aggregating Atomic Clocks for Time-Stamping.” Paper presented at 2020 IEEE International Conference on Service Oriented Systems Engineering (SOSE), Oxford, UK. August 03, 2020 - August 06, 2020.

of no more than 100 µs (Annex, Directive 2014/65/EU). Similar regulations from the U.S. Financial Industry Regulatory Authority have an even tighter clock synchronization requirement with no more than a 50 µs (SR FINRA 2016-005) divergence. General purpose computer clocks are not a good choice when accurate time is needed as they are known to have poor accuracy, with a possible drift of 5 s to 15 s per day (Lombardi, 2019). The Network Time Protocol (NTP) and atomic clocks are routinely used to synchronize actions and provide more accurate time for internet applications. However, even NTP and atomic clocks are not precise enough for some applications such as those that use blockchain technology where timestamp verification and transaction order is critical (Stavrou & Voas, 2017). In addition, a timestamp that is trusted and verifiable is needed to accurately sequence the blocks for blockchain consensus protocol (Kuhn et al., 2019). The purpose of this project is to propose a general purpose system called Timestamping Authority (TsA), which would be able to provide reliable, real-time, high precision timestamps preserving the order of events regardless of their number and time when they were processed. In this paper we will discuss the TsA motivation and design with blockchain technology as a use case.

2.0 Background

The most accurate and precise time is kept by atomic clocks, which measure the frequency of oscillations of atoms that happens to be very close to a constant number. For most applications, *accuracy* refers to the closeness of measurements to a particular true value (such as a physical location), while *precision* is the closeness of measurements to each other. UTC, as its name implies, involves *coordination* among a set of cooperating systems. In the context of timekeeping, accuracy refers to traceability to UTC, while precision is the degree of synchronization among a set of clocks.

As noted previously, financial regulators require highly accurate timestamps, with UTC traceability. Within financial systems, continuous monitoring of clock synchronization stability is required, with realtime comparisons to UTC references. Global Positioning System (GPS) signals are typically used to provide UTC traceability to a national institution such as NIST, the Research Institutes of Sweden, or others. UTC combines time scales known as International Atomic Time (TAI) and Universal Time (UT1), which rely on a weighted average time from

450 atomic clocks in 70 nations (for TAI) and observations of rotations of the Earth (for UT1). To reduce potential vulnerability to GPS jamming or spoofing, financial institutions may employ specialized high-power encrypted signal transmission to protect systems providing UTC traceability. NTP is commonly used to provide synchronization of clocks within variable-latency packet switched networks. For NTP implementations, atomic clocks are connected to servers that deliver time over the network. These servers in turn are connected to other servers, and so on (Mills, 2003). This tree-like design (Figure 1) is needed to reduce the load on the top-level servers.

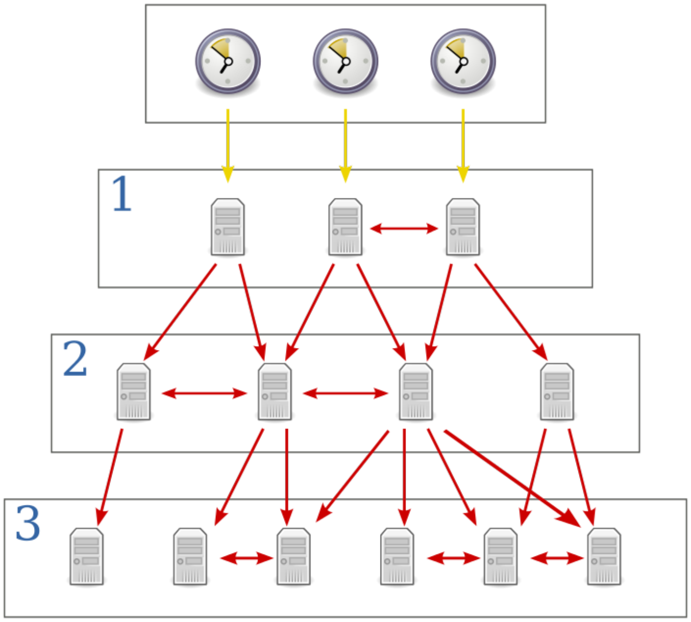


Figure 1: NTP Network Structure.

Yellow arrows indicate a direct connection; red arrows indicate a network connection.

This protocol operates over User Datagram Protocol (UDP), which is a transport layer network protocol. NTP version 3, used in the project, was standardized as Internet Engineering Task Force RFC-1305 (Mills, 1992), and is compatible with the latest version 4, RFC-7822 (Mizrahi & Mayer, 2016). The typical device might use one of the bottom-layer NTP servers to synchronize time. However, using the bottom layer can be insufficient in some situations, since some precision is lost on the way from the top. Improving precision becomes extremely expensive as the required level increases. The cause of this imprecision is due to the asymmetric network routes and network conditions (Mills, 2012).

A blockchain is an example of an application where timestamp accuracy is critical for verification and security. A blockchain is a series of timestamped immutable records that are managed by multiple computer entities. In order to add a new block to the chain, the candidate that wants to insert the block needs to perform a computation (e.g., calculate a hash of the concatenation of the previous block hash and a hash of the current data) and submit the result to the system for verification. The need for accurate timestamps comes into play when there are multiple candidates at the same time. In this case, only the candidate that finishes the hash calculation first will be added. All other candidates must start over with the new last block. This is a time consuming and expensive process. Figure 2 illustrates the addition of a timestamp using the basic methodology where only one candidate will succeed in being added to the chain and all others will have to start the hash calculation over – creating a type of race condition.

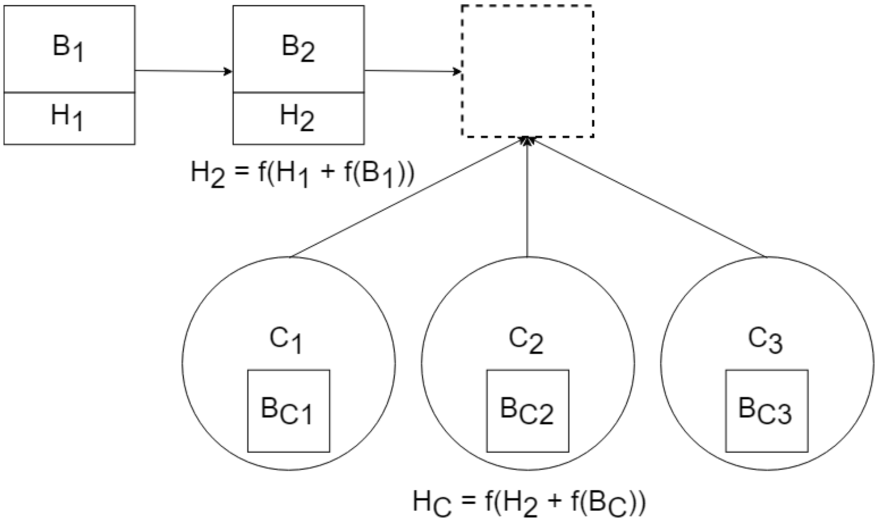


Figure 2: Adding a block to the blockchain with a basic timestamp (B – block; H – hash; f - hash function; C – candidate)

A possible solution to this problem is integrating a single TsA service into the system as shown in Figure 3.

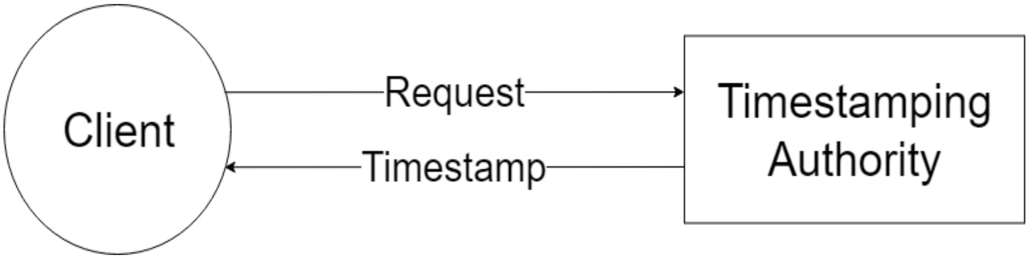


Figure 3: A Timestamping Authority Service

A TsA service could provide an accurate time on demand over a network. However, this does not resolve the situation where events happen simultaneously. As mentioned earlier, transaction order is critical in many applications. Figure 4 shows a single timestamping authority with the requests coming sequentially, thus keeping the accurate order of the requests. However, this design does not scale well and would work only for transactions that occurred geographically close to the TsA.

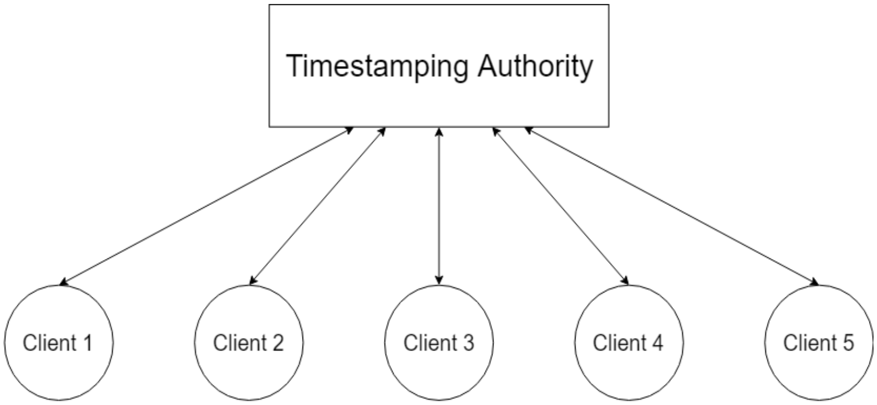


Figure 4: A single timestamping authority

A scalable timestamping solution would use several TsA servers, but then the question of time synchronization between them arises. An open specification called Chainpoint uses blockchain technology and the NIST randomness beacon to tackle this problem. A Chainpoint client stores a hash of the NIST randomness beacon and the timestamp on the blockchain:*Chainpoint (Timestamp + hash(random beacon))* → *blockchain*

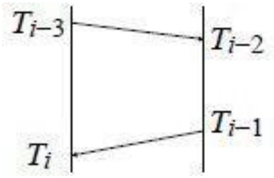
The NIST randomness beacon (beacon.nist.gov) produces a purely random 512-bit string every minute. Every random value is then stored in a blockchain and can be verified at any time.

Inserting the random number facilitates proving the authenticity of the timestamp at any minute on the timeline. This approach will both guarantee transaction order and provide verifiability to the timestamps. However, the time resolution of Chainpoint timestamps depends on the randomness beacon, which is one minute, so we need to look for a different approach.

3.0 TsA Design

To ensure accurate time and order, when a time request is made, the timestamp will be determined by using an aggregation from several atomic clocks. Specifically, the atomic clocks that are used as time sources for this project have publicly available NTP servers, which are synchronized to the clocks, thus providing precise time to the clients (see Sect. 3.1 for details of this process). The communication with the client also works over NTP. NTP can provide up to 1 ms precision under ideal network conditions, and the precision deteriorates as asymmetric network routes are introduced (see RFC-1305, Appendix F). When a network route is asymmetric, the time for a request to get to the server is different from the time to get the response. This situation usually happens when long-distance requests are made or when the network condition is unsatisfactory. Since only the nearest atomic clocks are used, to provide high precision, NTP also accounts for latency automatically (see RFC-1305). Figure 5 shows the transition of 4 timepoints in the UDP message that allows the calculation of correct network delay and local clock offset from the server clock.

The testing procedure consists of using one PC, running the Timestamping Authority Server locally, and running from one to three clients also locally. Due to everything running on the same PC, time for the request to get from the client to the server and from the server to the client is negligible, so the precision of the aggregation procedure can be tested. The server starts up, gets time from three atomic clocks, and the clients are run when the server is ready to respond, (i.e. it has calculated an average time from all three clocks). Then, since we may assume that network delays are not an issue, we can collect all the timestamps received by the clients and compare them to the order they were sent. The tests showed that the correct order is preserved when the requests are spaced out by 10 ms or more.



$$\delta = (T_i - T_{i-3}) - (T_{i-1} - T_{i-2})$$
$$\theta = \frac{(T_{i-2} - T_{i-3}) + (T_{i-1} - T_i)}{2}$$

Figure 5: Network Delay calculation from RFC-1305

Figure 6 shows an example time request from Event X. The TsA is retrieving the time from several atomic clocks and returns the precise time to the client.

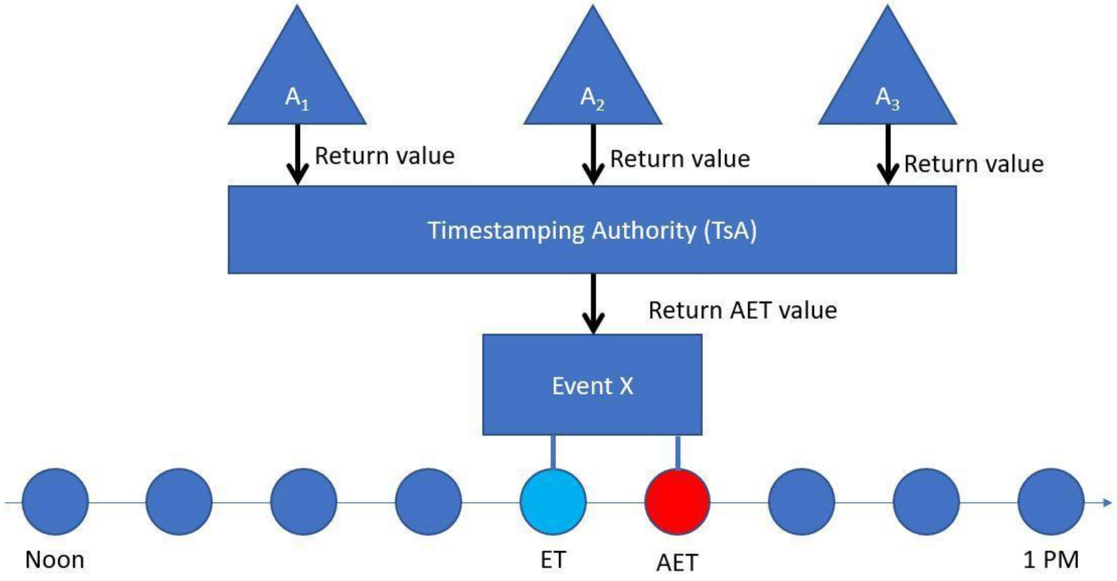


Figure 6: Scheme of TsA operation.
(ET - exact time of the event; AET - approximate exact time; A - atomic clock)

The timestamps are assigned to the candidates and form a queue, giving some time to the first-comer and notifying all other process candidates to wait (Figure 7). That is, the queue is formed among the candidates and each one is given some time to do the hash computation. If the client fails, the opportunity is given to the next candidate. This approach could reduce CPU time spent by the candidates, because they now have an opportunity to wait for the first arrival to finish before continuing their own computations. Timestamps will serve as a proof that the

Saidkhodjaev, Temur; Voas, Jeff; Kuhn, D. Richard; DeFranco, Joanna; Laplante, Phil. "Aggregating Atomic Clocks for Time-Stamping." Paper presented at 2020 IEEE International Conference on Service Oriented Systems Engineering (SOSE), Oxford, UK. August 03, 2020 - August 06, 2020.

candidate was indeed first to arrive. This approach would be especially useful if there are multiple servers accepting blocks and they need to be coordinated. In this case, they could query the TsA and receive timestamps that could be used across the servers.

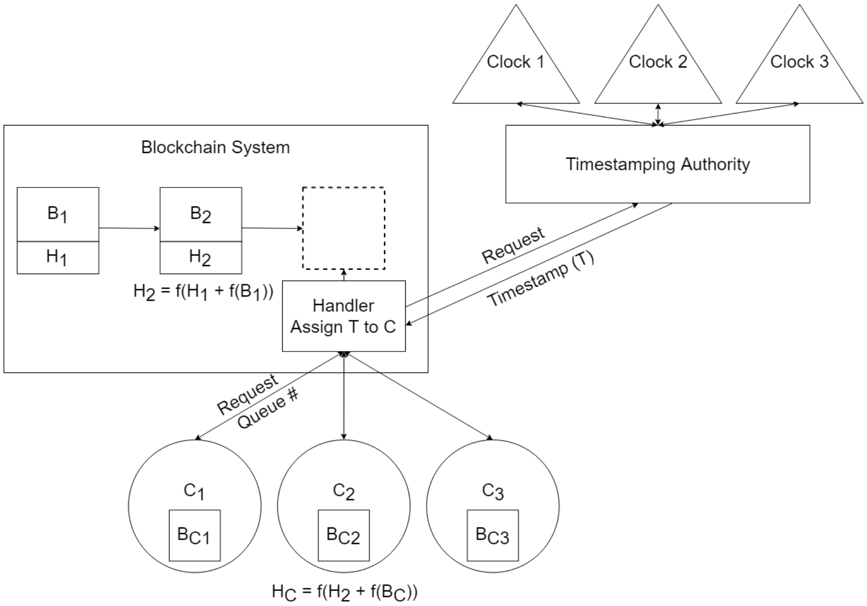


Figure 7: Adding a block to the blockchain with a TsA.

(B – block; H – hash; f - hash function; C – candidate)

It is important to note that the clients are not permitted to use local time due to possible imprecision. Maintaining accurate UTC time is difficult, and TsA does not claim that the time provided is accurate. Instead, the Time Stamping Authority establishes a common timescale for the system that would preserve order of timestamped events. In other words, Time Stamping Authority is precise, but not necessarily accurate. The project described in this paper was designed to demonstrate the effectiveness of an authority by aggregating time from several official reliable atomic clocks and use it to produce the common timescale. The prototype currently aggregates time from three nearby atomic clocks and computes the average of their times. (This is currently implemented as simple averaging, but some form of weighted average may be used in the future.) More effective formulae can be used after the aggregation. Tests with local server and client show that using this approach preserves the order of timestamps.

3.1 Clocks Used and Their Reliability

The International Bureau of Weights and Measures (IBWM) or Bureau International des Poids et Mesures (BIPM) in French, the organization that defines the International System of Units (SI) units, also provides the UTC time standard (https://www.bipm.org). This standard is created based on over 400 participating laboratories measuring time with their atomic clocks and reporting the measured time every 5 days to IBWM, which then calculates the weighted average based on the clock precision and produces UTC. The problem is that it is difficult to transmit clock data in real-time without losing accuracy, hence all calculated time points are in the past. Nevertheless, the data from the atomic clocks spread across the globe suggests that most of them are synchronized up to hundreds of nanoseconds (ns) (https://www.bipm.org/en/bipm-services/timescales/time-ftp/Circular-T.html), which is more than enough for the problem at hand. Within the U.S., the difference between UTC and the two commonly used clocks (NIST and Naval Observatory) is only a few ns. IBWM also publishes the Annual Report on Time Activities (https://www.bipm.org/en/bipm-services/timescales/time-ftp/annual-reports.html), which lists laboratories participating in creation of the time standard, and are known to be reliable. Some of those laboratories have publicly available NTP servers linked to the clocks, which are used in this project as time sources. For the North American region, possible clock servers are the National Institute of Standards and Technology NTP (https://tf.nist.gov/tf-cgi/servers.cgi), the United States Naval Observatory NTP (https://tycho.usno.navy.mil/NTP/) and the National Research Council of Canada NTP (https://nrc.canada.ca/en/certifications-evaluations-standards/canadas-official-time/network-time-protocol-ntp). One problem with using these clocks is that the public servers are vulnerable to distributed denial of service (DDoS) attacks, and therefore all clients making too frequent requests to them are banned (e.g., the NIST clock encourages no more than one request every 4 s). This problem has been solved by maintaining an offset from local time to each of the used clocks, and updating it every 10 s (this interval can be configured). Per Lombardi's estimations, most hardware clocks gain or lose about 5 s to 15 s per day (Lombardi, 2019), so with simple calculations, we see that the hardware clocks can gain or lose at most 2 ms every 10 s. While not perfect, the period of 10 s was chosen as a trade-off to both have acceptable precision, and not get the server banned by the atomic clock NTP servers. Therefore, even if a local clock is imprecise, it is very unlikely that any significant deviation will happen in 10 s.

4.0 Implementation Description

The NTP utilities library from Apache Commons Net was used in the project (https://commons.apache.org/proper/commons-net/). Some code was modified and everything else was used as a library. The project consists of four executables written in Java. The first and primary executable is the server, which can be configured and started via command line, specifying the NTP servers to be used at run time. The server periodically updates local time offsets of the used clocks and services client requests. There is a class *TimeStampingAuthorityServerRunner* with a main method that is compiled to the executable server. There are also two other classes that can be used as library code. One of these (*TimeStampingAuthority*) simply requests time from the clocks, maintains clock offsets and can provide aggregate time via the application programming interface (API), and another one (*TimeStampingAuthorityServer*) also runs an NTP server on a specified port. There is a separate Clock class that contains logic for maintaining the clock offsets and network communication with the atomic clock servers.

The second executable is the client for testing. It can be configured and started via command line, and it logs the time points received from the server to a .txt file. The client makes requests to the server at the specified address with a given frequency. Tests show that an interval of 10 ms or more should preserve events ordering on the TsA side. Even though such precision might not be enough for some applications, it has a benefit of simplicity: no need for advanced equipment or dedicated connections, just one program on any PC. Moreover, we would note that this is a proof-of-concept, so future developments may achieve higher precision.

The third executable takes the log file produced by the client and checks if timestamps the client received are in the ascending order. It can check multiple files at the same time and is also usable from a command line. The fourth executable takes several log files and combines them. This executable was used when several clients were run simultaneously to test how well TsA preserves order of events with its timestamps. All the executables are configurable. VT uses the Maven framework for dependency management and build automation. It also uses Git for tracking changes (https://github.com/usnistgov/blockmatrix/tree/master/TimeStampingAuthority).

Executables can be used on their own, but it is suggested that an interactive development environment (IDE) supporting Maven and Git is used in development, e.g. IntelliJ IDEA or Eclipse (both are free). There are four Maven modules corresponding to four executables in the project. 3 contain just one class with a main method, and one contains several classes mentioned above. There is a directory called "testing" which contains all the executables, a file with the clock data and a script for testing. Maven is configured in such a way that when the "package" command is run, all the modules are assembled into .jar files. When the "verify" command is run after that, the .jar files are moved into the correct directories for convenience. The "out" directory contains all 4 executables bundled with their dependencies as well as a library .jar, containing non-executable server-side classes without dependencies. This library jar can be used in other projects as a dependency, assuming that the Apache Commons Net library is included.

4.1 Installation

This project uses Java Development Kit (JDK) 12.0.1 (the current version as of this writing), but is tested to be backward-compatible with JDK 1.8.0_21.

If used as a .jar, the executables do not have any dependencies except for the Java runtime environment.

If used as Java classes, the Apache Commons Net library needs to be downloaded, which provides implementation of many network protocols, including NTP, along with utilities that make it easier to use these protocols in the program. You can include this library as a dependency in a Maven project using the information on

https://mvnrepository.com/artifact/commons-net/commons-net/3.6 or download the libraries as source code or as binaries on

https://commons.apache.org/proper/commons-net/download_net.cgi. The libraries are open source and are available under the Apache License, Version 2.0.

It is suggested that you use an IDE with support of Maven and Git if you would like to use them in the development of the project, because Maven downloads the dependencies for you and makes building and packaging easier and Git helps revert any changes and keep track of the history of the project. If not, a simple solution would be to rip out the classes of all 4 modules and put them in a new project together. Do not forget to add the Apache Commons Net library as a dependency.

5.0 Limitations

The order-preserving property of the current TsA implementation has been tested only locally. Namely, the procedure was to run the Time Stamping Authority server locally, together with one or more clients that would request timestamps from the server. Since the time on the PC is the same, more tests are needed to determine the reliability of TsA. Namely, clients could be located far from the server, and have their own reliable time source to compare the timestamps with. Based on this kind of testing, the conclusion about order-preservation can be made.

Sometimes several subsequent timestamps get "merged" and have the same value, even though it is known that the events were supposed to have a time interval between them. For example, there could be several timestamps with values 1, 1, 1, 4, 5, while the events were supposed to happen at times 1, 2, 3, 4, 5. It is likely that this happens due to low computing power of the development machine used, and as a consequence of running several processes for actual testing of the program and other processes running in the background. If the events are spaced out with the stated time difference - 10 ms, then the weak order is always preserved. By weak order we mean the sequence of numbers where two numbers can be equal, not strictly increasing.

Implementation factors are a consideration in this merging phenomenon. The prototype was implemented in Java, which sometimes has unpredictable runtime due to garbage collection, and it was not the only program running on the PC, so the operating system (OS) could take away processor time for a fraction of a second, corrupting the timestamps. It is expected that if the implementation had adequate performance and priority is given to the time-stamping server, these anomalies should not be present, and the resulting timestamps should form strictly increasing sequence. Additional testing will be used to evaluate this condition.

1. NTP assumes that the offset from the client clock is to be calculated by the client, so the TsA has no way to know if the client completed the timestamp calculation correctly.
2. When requesting time from the clocks and serving time to the client, only one NTP-request is made, since acceptably low latency network conditions are assumed. It is suggested to make several NTP-requests and take the one with the lowest round trip time to exclude asymmetric route errors.
3. RFC-1305 suggests a formula for combining several clock times to increase accuracy and precision. This could be used instead of a simple average of all the clocks, which is currently used.

6.0 Conclusion

In the real estate community, it is said that property value is based on “location, location, location.” But we also know that timing is everything. In sports, the end of a game can be decided in a single second. In financial transactions, it comes down to milli- and microseconds. Here, we have shown that atomic clock aggregation is possible and that it works better if the clocks are somewhat geographically co-located due to latency. We have argued that a Timestamping Authority (TsA) is a feasible approach to creating timestamps of higher integrity.

We plan to continue this research to better understand the impact of latencies on the accuracy of the aggregated clock results.

7.0 References

Lombardi, M., “Computer Time Synchronization,” Time and Frequency Division, National Institute of Standards and Technology, https://tf.nist.gov/service/pdf/computertime.pdf, retrieved, 8/17/19.

Mills, D., “Network Time Protocol (Version 3) Specification, Implementation and Analysis,” RFC 1305, March 1992, https://tools.ietf.org/pdf/rfc1305.pdf, retrieved 8/17/19.

Mizrahi, T., Mayer, M., “Network Time Protocol Version 4 (NTPv4) Extension Fields,” Internet Engineering Task Force (IETF), March 2016, https://tools.ietf.org/pdf/rfc7822.pdf, retrieved 8/17/19.

Kuhn, R., Yaga, D., Voas, J., “Rethinking Distributed Ledger Technology,” Computer, Feb 2019, pp. 68-72.
Stavrou, A., Voas, J., “Verified Time,” Computer, March 2017, pp. 78-82.

Directive 2014/65/EU of the European Parliament and of the Council with regard to regulatory technical standards for the level of accuracy of business clocks
http://ec.europa.eu/finance/securities/docs/isd/mifid/rts/160607-rts-25_en.pdf
Annex: http://ec.europa.eu/finance/securities/docs/isd/mifid/rts/160607-rts-25-annex_en.pdf

SR-FINRA-2016-005. Proposed Rule Change to Reduce the Synchronization Tolerance for Computer Clocks that are Used to Record Events in NMS Securities and OTC Equity Securities
https://www.finra.org/industry/rule-filings/sr-finra-2016-005

Mills, D.L. (2003). A brief history of NTP time: memoirs of an Internet timekeeper. Computer Communication Review, 33, 9-21.

Mills, D.L. (2012). Executive Summary: Computer Network Time Synchronization. https://www.eecis.udel.edu/~mills/exec.html.

Multiple Changepoint Analysis of Noisy Nonlinear Data with an Application to Modeling Crack Growth in Additively Manufactured Titanium

Lucas Koepke[*][†] Jolene Splett[‡] Timothy Quinn[‡] Nikolas Hrabe[‡]
Jake Benzing[‡] Michael Frey[‡]

Abstract
Noisy measurement data pose a challenge for changepoint analysis, especially in the presence of multiple changepoints and when the model is nonlinear. We explore various approaches to estimating changepoints and their standard errors under these conditions. We consider whether adding a monotonicity constraint improves the changepoint estimates and reduces their standard errors. We finish with a novel application to material science using crack growth data from additively manufactured titanium. As cyclic loading is applied to a test specimen, crack growth can be partitioned into three regimes: slow-growth, mid-growth, and high-growth. We improve estimates of the transition points between these regimes versus those made by experts in the field by adding confidence bounds to the changepoint locations, allowing for designed experiments to study treatment effects on changepoint location.

Key Words: changepoint, isotonic regression, nonlinear least squares, pool-adjacent-violators algorithm

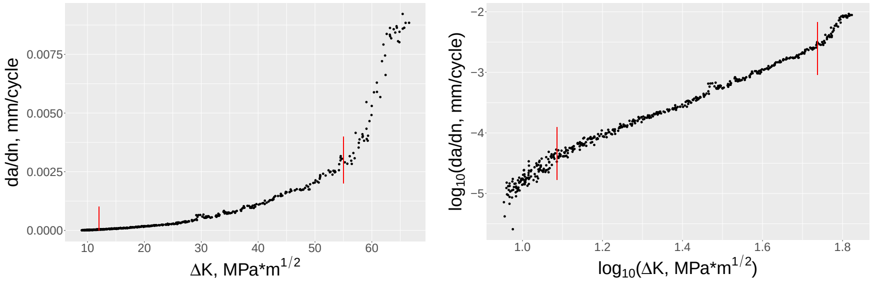
1. Introduction

As a material is subjected to cyclic fatigue loading, cracks can form and grow over time even when the maximum applied force is well below the yield strength of the material. To measure the resistance of a material to fatigue crack propagation, a notched sample is cyclically loaded for many thousands of cycles. For each force cycle, the crack size a is estimated and the cyclic force range, $\Delta P = P_{max} - P_{min}$ is recorded. For analysis, the stress intensity factor, ΔK, is calculated from ΔP, a, and the geometry of the specimen, since during fatigue testing the force range is constant, while ΔK increases as the crack propagates. The stress intensity can be calculated for any part where the forces and geometry are known. ΔK is plotted against the change in crack length per cycle or fatigue crack rate, $\frac{da}{dn}$ (see Figure 1).

Figure 1a shows data on the linear scale. For lower values of ΔK, the crack shows minimal growth. This transitions to an elbow region as ΔK increases, where the crack growth rate can be predicted with a power law (Hertzberg, 1996, p. 614). This in turn leads to a short period of rapid crack growth. The transitions between these three regimes (slow-growth, mid-growth, and high-growth) can be used to characterize the fatigue properties of the material. Because the transitions are difficult to identify on the linear scale, both $\frac{da}{dn}$ and ΔK are typically transformed by taking the log (base 10) (Figure 1b). The transformation changes the elbow region into a linear region, with a lower changepoint occurring at the transition into the linear region and an upper changepoint at the transition out of the linear

[*] Associate, National Institute of Standards and Techonology, Boulder, CO 80305
[†] University of Colorado, Department of Physics, Boulder, CO 80309
[‡] National Institute of Standards and Technology, Boulder, CO 80305

Koepke, Lucas; Splett, Jolene D.; Quinn, Timothy; Hrabe, Nik; Benzing, Jake; Frey, Michael. "Multiple Changepoint Analysis of Noisy Nonlinear Data with an Application to Modeling Crack Growth in Additively Manufactured Titanium." Paper presented at 2019 Joint Statistical Meetings, Denver, CO, US. July 28, 2019 - August 01, 2019.



(a) Original data.

(b) Log-transformed data.

Figure 1: Example of fatigue crack growth data from additively manufactured titanium on the linear scale (left) and after taking the log (base 10) of both the stress intensity factor ΔK and the crack growth rate $\frac{da}{dn}$ (right). Possible changepoint locations are shown by the vertical red lines.

region. These changepoints are typically determined through inspection by experts in the field. Our challenge is to estimate these two changepoints objectively.

We proceed as follows. Since we don't know the true changepoints for the experimental data, we start with a simulation study both to compare methods using data with a known changepoint and to explore how aspects of the data might bias the estimates. Additionally, we investigate whether it is possible to reduce noise in the data knowing that crack growth should always increase as ΔK increases. We finish by estimating the changepoints on experimental data and comparing results with changepoints estimated using inspection by expert engineers. All computation was performed using R (R Core Team, 2018).

2. Methods

We assume that the model should be linear between the changepoints, so we concentrate on methods that can be made to accommodate this constraint. Specifically, we focus on fitting models using nonlinear least squares (NLS). We can parameterize the function that we fit to approximate different shapes below, between, and above the two changepoints. In each case, the function segments are constrained to meet at the changepoints, and the changepoint locations are estimated as parameters. The approach also has the advantage of providing standard errors on the parameter estimates and model fit diagnostics. For example, the model for estimating one changepoint with one linear and one quadratic segment is

$$\begin{aligned}\hat{y} =&I_{x\leq \text{changepoint}}(a_1 + b_1 * x)+ \\ &I_{x>\text{changepoint}}(b_2 * x + c_2 * x^2)\end{aligned} \tag{1}$$

If c_2 is set to zero this model reduces to one with two linear segments meeting at the changepoint.

2.1 Noise reduction

We know that $\frac{da}{dn}$ should be monotonically increasing with ΔK (cracks do not decrease in size in engineering materials under cyclic force), so it is natural to consider the benefit of isotonic regression (Barlow and Brunk, 1972; Dykstra, 1981; Robertson et al., 1988). Isotonic regression is unique in that it is guaranteed to reduce

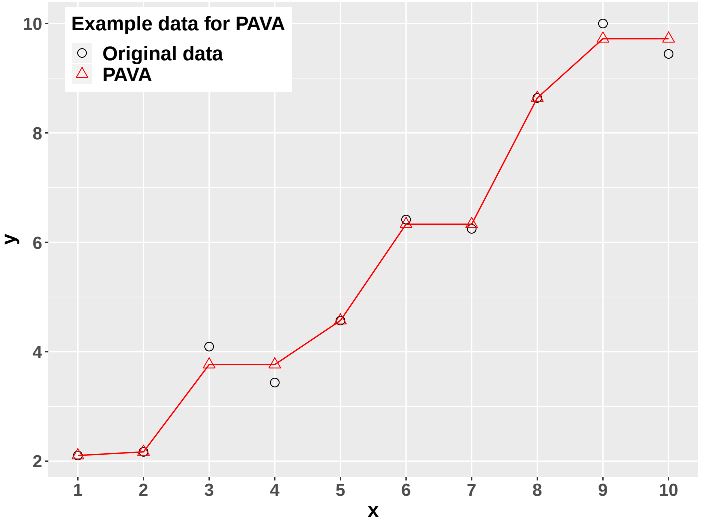


Figure 2: Example of PAVA using simulated data. Black circles are the original data points, the red triangles show the result after PAVA. If the original data is already increasing, PAVA does nothing, but the pooling is evident whenever a y value decreases.

mean-squared error when estimating a monotonic mean function. We consider the isotonic regression calculated using the pool-adjacent-violators algorithm (PAVA) (Ayer et al., 1955). PAVA is implemented in R in the `isotone` package (Mair et al., 2009).

PAVA works as follows. Start with the data (x_i, y_i) for observations $i \in 1, \ldots, n$ ordered such that $x_i \leq x_{i+1}$. At this point the y_i's have not been touched. We assume equal weights. Start at $i = 1$ and check whether y_2 is less than y_1. If this is the case, the monotonicity constraint is violated, so pool y_1 and y_2 by replacing both with their average $\frac{y_1+y_2}{2}$. Then proceed as follows for $i \in 2, \ldots, n-1$:

1. Check for a monotonicity violation (y_{i+1} is less than y_i). If that is the case, pool y_{i+1} and y_i.

2. If the new value of y_i is now less than y_{i-1}, pool all three (y_{i-1}, y_i, and y_{i+1}) by replacing the three values with their average. Then for $j \in i-1, \ldots, 1$ repeat until $j = 1$ or $y_j < y_{j+1}$.

3. Proceed to the next i.

After the PAVA computation has completed, the y's are monotonically non-decreasing (Figure 2). Even though PAVA appears to naturally apply to fatigue crack growth data due to the monotonicity, it is not clear whether PAVA will bias the changepoint estimates, or to what extent the associated standard errors will be affected.

3. Simulation study

The simulation study is set up to explore different situations that might arise in the experimental data in a controlled way. We simulate data using (1) for the one

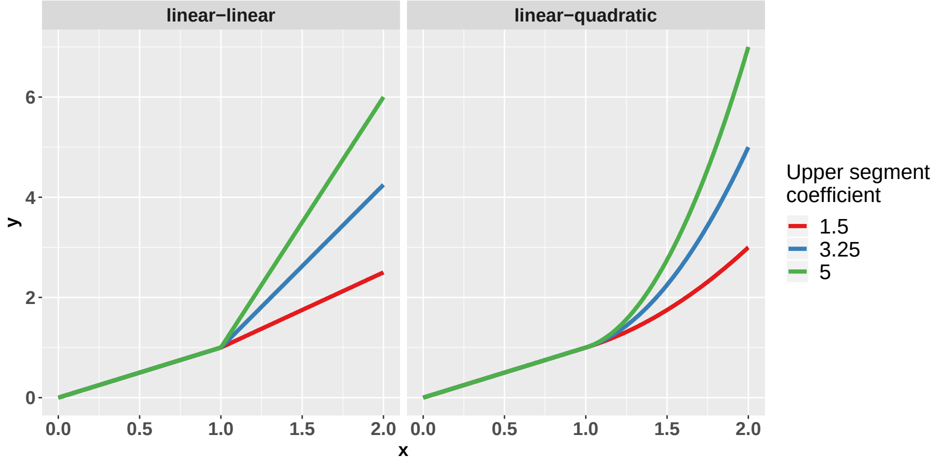


Figure 3: Data generation models shown without noise. Both have a linear first section with $b_2 = 1$ up to the changepoint at $x = 1$. Multiple parameter values for the upper segments are shown, corresponding to the slope on the upper segment (b_2) in the linear-linear model and the coefficient on the quadratic term (c_2) for the linear-quadratic model.

changepoint case since including a second changepoint is a straightforward extension of the NLS function. The first segment is a line with b_1=1 for x between 0 and 1. The changepoint is at $x = 1$. For x between 1 and 2, we consider both a linear and a quadratic segment, the quadratic is chosen because it represents the simplest nonlinear case.

Two parameters are varied when simulating data. The first is the amount of additive noise. In the experimental data, the error relative to the range of the y measurements is about 1 %, so we consider a range of additive noise values including: 1 %, 2 %, 3 %, and 5 %. The other parameter we vary is the shape of the second segment after the changepoint. In the linear-linear model where c_2=0, this parameter is the slope of the second segment (b_2), and will range between 1.5 and 5. For the linear-quadratic model, we fix the slope of the quadratic segment to be 1 at the changepoint (b_2=1) so the function is continuous and smooth, but we vary the coefficient on the quadratic term (c_2) between 0.5 and 5. Changing c_2 is equivalent to changing the second derivative (differing by a factor of 2). Examples of these data generation models with different parameter values are shown in Figure 3.

For each model, parameter value, and relative noise level, we simulate 10,000 data sets, each with 200 points below the changepoint and 200 points above, and fit linear-linear and linear-quadratic models using NLS. The output of each iteration is an estimated changepoint location, and we use these to compute the bias and standard error of the changepoints. The second part of the simulation will treat noise reduction using PAVA. Although PAVA should reduce the noise in the data with no penalty in overall model fit, it is not clear whether PAVA will introduce bias in the specific parameter estimate for the changepoint. Thus for each simulated data set, we also fit the model to PAVA smoothed data, and compare the bias and standard error of the changepoint parameter estimates for unsmoothed and PAVA-smoothed data.

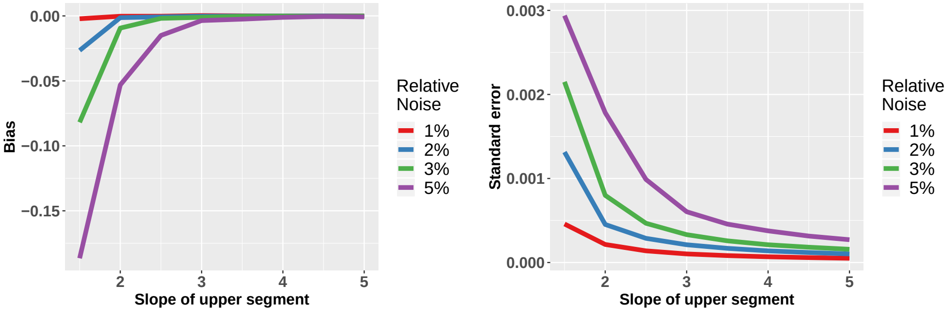


Figure 4: Bias (left) and standard error (right) for the estimated changepoint obtained by fitting the linear-linear model to linear-linear data versus the value of b_2 used to simulate the data.

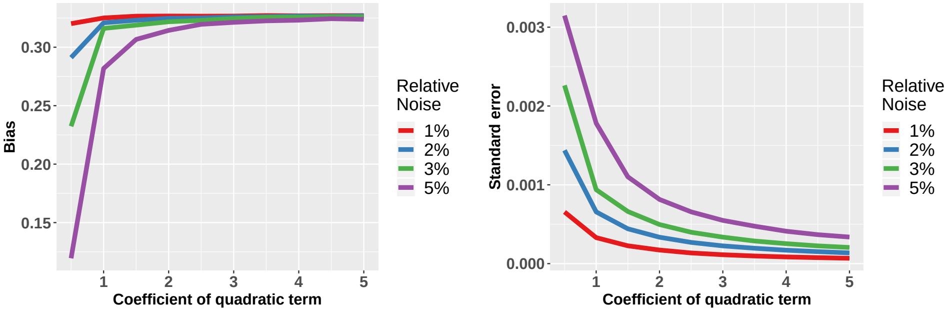


Figure 5: Bias (left) and standard error (right) for the estimated changepoint obtained by fitting the linear-linear model to linear-quadratic data versus the value of c_2 used to simulate the data.

3.1 Bias and standard error

Figure 4 shows the bias (left) and standard error (right) of the estimated changepoint for the linear-linear model fit to the linear-linear data. On the x-axis is the slope of the upper segment, ranging between 1.5 and 5. In the noisiest scenario (5 % relative noise) with the least slope difference ($b_2 = 1.5$), the bias is -0.18. When $b_2 = 2$, the bias drops to just -0.05 and keeps improving as the slope increases. The standard error follows a similar pattern.

Results for bias and standard error of the estimated changepoint for the linear-linear model fit to linear-quadratic data are shown in Figure 5. The shape of the curves for bias and standard error are similar to the previous case (fitting a linear-linear model to linear-linear data) but now instead of the bias trending towards zero it settles at roughly 0.32. The standard error decreases for all relative noise levels as the coefficient on the quadratic term increases.

Results for bias and standard error of the estimated changepoint for the linear-quadratic model fit to linear-quadratic data are shown in Figure 6. When c_2 is between 0.5 and 1.5, the biases are near zero but the standard errors at the four relative noise levels are highest at the 0.5 value, decreasing as c_2 increases. The bias is negative for coefficients 2 through 5. A negative bias here indicates that the transition to the quadratic segment, the changepoint, is estimated to be too low, so the quadratic segment consistently takes over part of the linear segment.

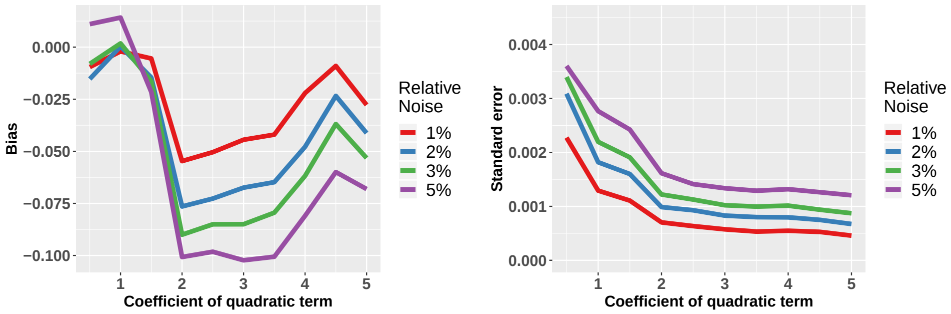


Figure 6: Bias (left) and standard error (right) for the estimated changepoint obtained by fitting the linear-quadratic model to linear-quadratic data versus the value of c_2 used to simulate the data.

The results of the simulation show that fitting a linear-linear model to linear-linear data performs well. The recommendation is not as clear for the linear-quadratic data, since fitting a linear-quadratic model does not perform consistently. The parameter combinations with biases near zero have the largest standard errors, and there are erratic jumps in the bias for some coefficient values. Fitting a linear-linear model to the linear-quadratic data shows higher bias in the changepoint estimate, but the results are consistent over different coefficient values, and the standard errors are consistently lower compared with the linear-quadratic fit.

3.2 PAVA results

In the simulation study, we explore two ways that PAVA can influence the analysis. The first is to show the improvement in model fit, and the second is to explore the effect on the changepoint estimates themselves.

In terms of model fit, the reduction in mean-squared error from unsmoothed to PAVA-smoothed data is shown in Figure 7. The percent change was calculated going from unsmoothed to PAVA-smoothed data, relative to the unsmoothed data. For both cases where the model fit matches the data model, PAVA reduced the mean-squared error of the NLS fit by over 80 %. This reduction is larger at high relative noise levels, and decreases as the upper segment becomes steeper. The third case, fitting a linear-linear model to linear-quadratic data, shows similar improvement (over 80 %) when b_2 is 0.5 or 1, but offers less of an improvement as b_2 increases. When $b_2 = 5$, PAVA reduced the mean-squared error by less than 40 % at the lowest relative noise level.

The effect of PAVA on the bias of the changepoint estimates is shown in Figure 8. This plot shows the difference in absolute values, $|\text{bias}_{\text{unsmoothed}}| - |\text{bias}_{\text{PAVA}}|$. A positive value means that PAVA reduces the bias, while a negative value means the opposite. We chose to use this difference instead of calculating a percentage change since many values are close to zero, leading to unstable relative changes.

For a linear-linear model fit to linear-linear data, PAVA reduces the magnitude of the bias when $b_2 = 1.5$ for the three highest relative noise levels (2 %, 3 %, and 5 %). When $b_2 = 2$, PAVA reduces the magnitude of the bias only for the 5 % relative noise case. PAVA increases the bias for the other combinations of slope and relative noise, but only by 0.01 or less.

PAVA does not improve the bias when fitting a linear-linear model to linear-

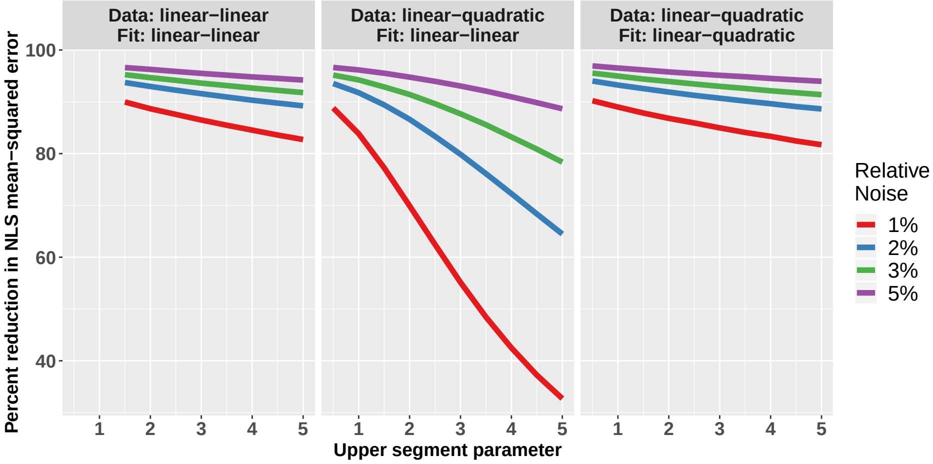


Figure 7: Percent change in mean-squared error of the NLS fit going from unsmoothed to PAVA smoothed data. (left) linear-linear model fit to linear-linear data, (center) linear-linear model fit to linear-quadratic data, and (right) linear-quadratic model fit to linear-quadratic data.

quadratic data. Bias is actually increased for all combinations of relative noise and the coefficient on the quadratic term (c_2), although this increase is 0.02 or less for relative noise of 1 %. The difference in bias decreases for all relative noise levels as c_2 increases.

When fitting a linear-quadratic model to linear-quadratic data, PAVA reduces the bias at all noise levels for c_2 values from 1.5 to 4, although by less than 0.025 in all cases. Bias increases for a quadratic coefficient of 1, but by less than 0.01.

While PAVA substantially reduces the overall error in the model fit, it does not offer a consistent reduction in bias for the changepoint estimate. However, in cases where PAVA increases the magnitude of the bias, the increase is less than 0.01 when the correct model is fit to the data. The bias increase is more than 0.02 when fitting the linear-linear model to linear-quadratic data when the c_2 is less than 2. Although PAVA does not consistently reduce the bias of the changepoint estimates, the reduction in mean-squared error from the model fit means that PAVA could be a useful step in estimating changepoints with NLS under the right conditions (for example, the linear-linear model fit to linear-linear data with high noise and a small change in slope). PAVA should not be used for the case where a linear-linear model is fit to linear-quadratic data.

4. Analysis of experimental data

We turn now to the problem of estimating the two changepoints on the experimental data. We extend the linear-linear and linear-quadratic models from the simulation study to now include two changepoints. For simplicity we call the linear-linear-linear model the “1-1-1” model and the quadratic-linear-quadratic model the “2-1-2” model, from the order of the polynomials in each segment. The equation that

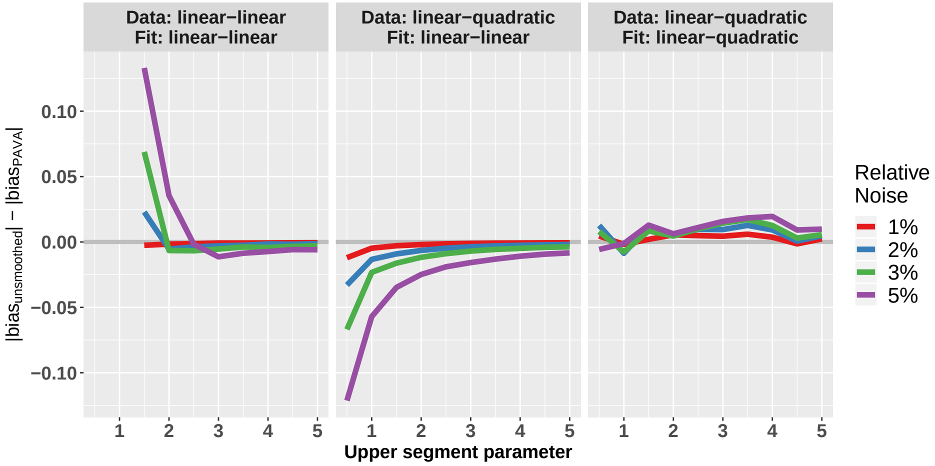


Figure 8: Difference in the absolute value of the bias for unsmoothed data and the absolute value of bias for PAVA-smoothed data. A positive value indicates that the magnitude of the bias is smaller after PAVA, while a negative value indicates the opposite.

Parameter	1-1-1 model	2-1-2 model	Expert
Lower changepoint	1.087 (0.00579)	1.152 (8.735)	1.06 (0.0107)
Upper changepoint	1.735 (0.00469)	1.667 (0.0283)	1.75 (0.00974)
b_0	5.180 (0.162)	2.710 (161.622)	N/A
c_0	N/A	-9.268 (1.756)	N/A
a_1	-7.333 (0.0252)	-7.299 (0.0368)	N/A
b_1	2.735 (0.0179)	2.712 (0.0261)	N/A
b_2	6.893 (0.353)	1.865 (0.783)	N/A
c_2	N/A	21.821 (3.954)	N/A

Table 1: Parameter estimates for the 1-1-1 and 2-1-2 models, and the mean values from the expert engineers. Standard errors for each point estimate are shown in parentheses. Units on the changepoint estimates are $\log_{10}(\text{MPa} * \text{m}^{1/2})$.

we fit using NLS is given by

$$\begin{aligned}\hat{y} =& I_{x<\text{lower changepoint}}(b_0 * x + c_0 * x^2) + \\ & I_{\text{lower changepoint}\leq x\leq \text{upper changepoint}}(a_1 + b_1 * x) + \\ & I_{x>\text{upper changepoint}}(b_2 * x + c_2 * x^2)\end{aligned} \tag{2}$$

where c_0 and c_2 are set at zero for the 1-1-1 model. There are no intercept terms estimated for the lower and upper segments because they are not free parameters. Since we do not know the true changepoints, we first had three expert engineers provide their best guess of the changepoint locations, independent of one another, for comparison with our results. Results for the 1-1-1 and 2-1-2 models are compared with the expert estimates in Table 1. This comparison illustrates why we use the 1-1-1 model for fitting even though the fatigue crack growth data appears to follow a 2-1-2 model. The 2-1-2 model fails to produce a lower changepoint with a reasonable standard error, and the 2-1-2 estimates are not as close to the expert changepoints as the 1-1-1 estimates. The experimental data may correspond to a high bias,

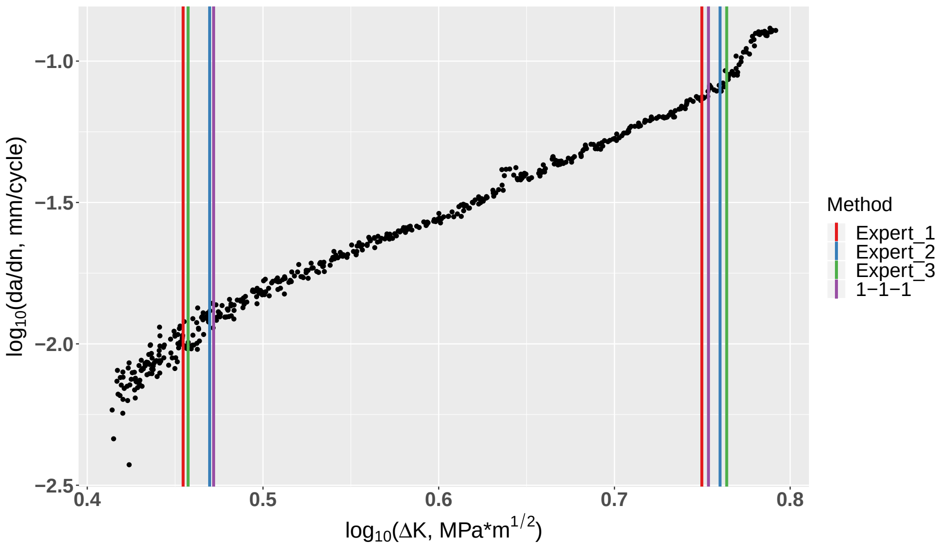


Figure 9: Experimental data with changepoints from each of the three experts and the 1-1-1 model. The 1-1-1 estimates agree almost exactly with at least one expert for both the upper and lower changepoints.

high standard error case in the simulation study where a linear-linear model is fit to linear quadratic data. In addition, the 2-1-2 model may not be a good fit to the experimental data. The 1-1-1 estimates are much closer to the average values from the experts for both changepoints, and the estimates have reasonable standard errors. Note that the 1-1-1 and 2-1-2 models estimate almost identical values for the slope and intercept of the central linear segment.

Individual estimates from the three experts and the 1-1-1 model are shown in Figure 9. The expert estimates show variability in the changepoint locations, but the estimates from the 1-1-1 model are both in the right area and very close to at least one expert for both the upper and lower changepoints. Results for data from other fatigue tests are similar, so we conclude that the 1-1-1 method is at least as reliable as an expert engineer in estimating the two changepoints.

For the experimental data, Figure 10 shows the point estimates and 95 % confidence intervals for three methods: experts (mean ± 2×standard error), the 1-1-1 model, and the 1-1-1 model with PAVA. While PAVA isn't recommended for fitting a linear-linear model to linear-quadratic data, we estimate changepoints for PAVA-transformed experimental data for completeness. Based on the simulation results for the case where a linear-linear model is fit to linear-quadratic data, the upper changepoint should be biased high; however, the 1-1-1 model produces an upper changepoint that is less than the average computed for the experts. Although the point estimates for the 1-1-1 method with and without PAVA are similar, PAVA does reduce the widths of the confidence intervals.

5. Conclusions

The simulation study provides insight into the behavior of changepoint estimation using NLS for a variety of scenarios. Fitting a linear-linear model to linear-linear data produces the most consistent changepoint estimates, with biases smaller than

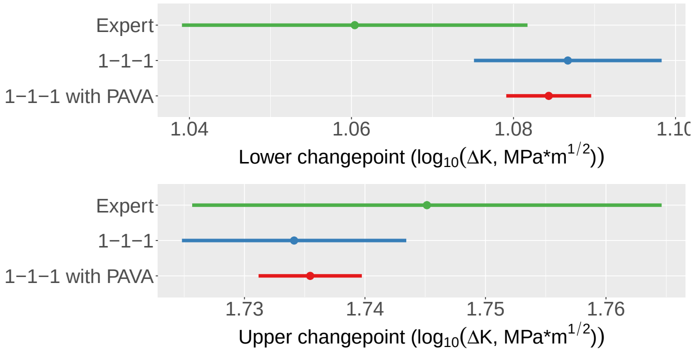


Figure 10: Comparison of point estimates and 95 % confidence intervals for the lower changepoint (top) and the upper changepoint (bottom).

0.01 in most cases. When fitting a linear-quadratic model to linear-quadratic data, many biases were in the range of -0.025 to -0.1. Changepoint estimates are biased, converging to a bias of 0.32, for all cases considered when fitting a linear-linear model to linear-quadratic data. Additionally, PAVA always reduces the mean-squared error for the model fit, but did not always reduce the bias in the changepoint estimates and should be used with caution.

The application to fatigue crack growth data shows that the 1-1-1 model provides realistic changepoint estimates even though the model doesn't accurately represent the data in the lower and upper segments. This method performs quite well compared with changepoint estimates made by experts in the field.

Future work will focus on two areas. First, we will estimate changepoints in the context of a designed experiment to evaluate how material conditions affect the changepoint locations. The second area of work involves missing data. Because valid data can only be collected while the crack is below a certain percentage of the sample width (ASTM E647), a particular test may only provide data on one of the two changepoints. Utilizing data from these tests would be useful, since fatigue tests are expensive.

This work builds the foundation for a methodical, data-driven approach to the analysis of fatigue crack growth data. The objectivity provided by our statistical approach will be useful to the scientific community.

References

M. Ayer, H.D. Brunk, G.M. Ewing, W.T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *The annals of mathematical statistics*, pages 641–647, 1955.

R.E. Barlow and H.D. Brunk. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147, 1972.

R.L. Dykstra. An isotonic regression algorithm. *Journal of Statistical Planning and Inference*, 5(4):355–363, 1981.

R.W. Hertzberg. *Deformation and fracture mechanics of engineering materials*. J. Wiley & Sons, 1996. ISBN 9780471012146.

P. Mair, K. Hornik, and J. de Leeuw. Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32(5):1–24, 2009.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

T. Robertson, F.T. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley, 1988. ISBN 9780471917878.

The Characteristics of a 1 m Methanol Pool Fire

Kunhyuk Sung[a], Jian Chen[a], Matthew Bundy,[a] and Anthony Hamins[a*]

[a] National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, Maryland, USA

* Tel: +1-301-975-6598, email: anthony.hamins@nist.gov

Highlights:

- The heat release rates determined by calorimetry and mass loss compared favorably
- Temperature profiles in the radial and axial directions were measured
- Gas temperatures were estimated considering radiative loss and thermal inertia effects
- The radiative fraction was calculated as 0.22 ± 16 % from heat flux measurements

Abstract:

A series of measurements was made to characterize the structure of a 1 m diameter methanol (CH_3OH) pool fire steadily burning with a constant lip height in a quiescent environment. Time-averaged local measurements of gas-phase temperature were conducted using 50 μm diameter, Type S, bare wires, with a bead that was approximately spherical with a diameter of about 150 μm. The thermocouple signals were corrected for radiative loss and thermal inertia effects. The mass burning rate was measured by monitoring the mass loss in the methanol reservoir feeding the liquid pool. The heat release rate was measured using oxygen consumption calorimetry. The heat flux was measured in the radial and vertical directions and the radiative fraction was estimated, which corresponded to previous results.

Keywords: Heat release rate; Temperature distribution; Burning rate; Heat flux distribution; Radiative fraction

1. Introduction

The focus of this study is to characterize the burning of a 1 m diameter pool fire steadily burning in a well-ventilated quiescent environment. Pool fires are a fundamental type of combustion phenomena in which the fuel surface is flat and horizontal, which provides a simple and well-defined configuration to test models and further the understanding of fire phenomena. In this study, methanol is selected as the fuel. Fires established by methanol are unusual as no carbonaceous soot is present or emitted. This creates a particularly useful testbed for fire models and their radiation sub models that consider emission by gaseous species - without the confounding effects of radiative exchange due to soot.

Many studies have been reported on the structure and characteristics of 30 cm diameter methanol pool fires, including the total mass loss rate [1-3], mean velocity [4], pulsation frequency [4] and gas-phase temperature field [4, 5]. With so many measurements

Sung, Kunhyuk; Chen, Jian; Bundy, Matthew; Hamins, Anthony. "The Characteristics of a 1 m Methanol Pool Fire." Paper presented at 13 International Symposium on Fire Safety Science (IAFSS 2020), Waterloo, CA. April 26, 2020 - May 01, 2020.

characterizing the 30 cm methanol pool fire, it is a suitable candidate for fire modeling validation studies [3, 6-8]. On the other hand, research on the detailed structure and dynamics of larger pool fires is limited. Tieszen, *et. al.* [9, 10] used particle imaging velocimetry to measure the mean velocity field in a series of 1 MW to 3 MW methane and hydrogen pool fires burning in a 1 m diameter burner. Klassen and Gore [11] reported on flame height and the heat flux distribution near 1.0 m diameter pool fires burning a number of fuels including methanol. They used the same burner as this study, but with a 5 mm (rather than 10 mm as used here) lip height. This study complements Ref. [11] by also measuring the local flame temperature throughout the flow field, the heat release rate using oxygen consumption calorimetry, and the radiative fraction determined by a single location measurement.

Use of fire modeling in fire protection engineering has increased dramatically during the last decade due to the development of practical computational fluid dynamics fire models and the decreased cost of computational power. Today, fire protection engineers use models like the Consolidated Fire and Smoke Transport Model (CFAST) and the Fire Dynamics Simulator (FDS) to design safer buildings, power plants, aircraft, trains, and marine vessels to name just a few types of applications [6, 12]. To be reliable, the models require validation, which involves a large collection of experimental measurements. An objective of this report is to provide data for use in fire model evaluation by the fire research community. Also, it is of interest to compare the burning characteristics of the 30 cm methanol pool fire with the results presented here for a 1 m diameter methanol pool fire.

2. Experimental Methods

Steady-state burning conditions were established before measurements were initiated. A warm-up period of 10 min was required for the mass burning rate to be steady. Since back diffusion of water slowly accumulates in the fuel pool in methanol fires, fresh fuel was used between experiments. The purity of the methanol was 99.99 % by mass and the density was 792.7 kg/m^3 at 20 °C, according to a report of analysis provided by the supplier. Experiments were conducted under an exhaust hood located 4 m above the burner rim. The effect of ambient convective currents on the fire were minimized by closing all inlet vents in the laboratory. The exhaust consisted of a large round duct (1.5 m diameter) located 6.0 m above the floor [13]. The smallest exhaust flow possible (about 4 kg/s) was used, helping to avoid perturbations (such as flame lean) and minimizing the influence of the exhaust on fire behavior. This led to the establishment of an unusually symmetric and recurring fire. The experiments were repeated three times.[*]

[*] Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

2.1. Pool Burner Setup

A circular steel pan with an inner diameter (*D*) of 1.00 m, a depth of 0.15 m, and a wall thickness of 0.0016 m held the liquid methanol. An image of the burner is seen in Fig. 1. The bottom of the burner was water cooled. The burner was mounted on cinder blocks such that the burner rim was about 0.3 m above the floor. A fuel overflow basin included for safety extended 3 cm beyond the burner wall at its base. The fuel inlet was insulated and covered with a reflective foil to prevent preheating of the fuel.

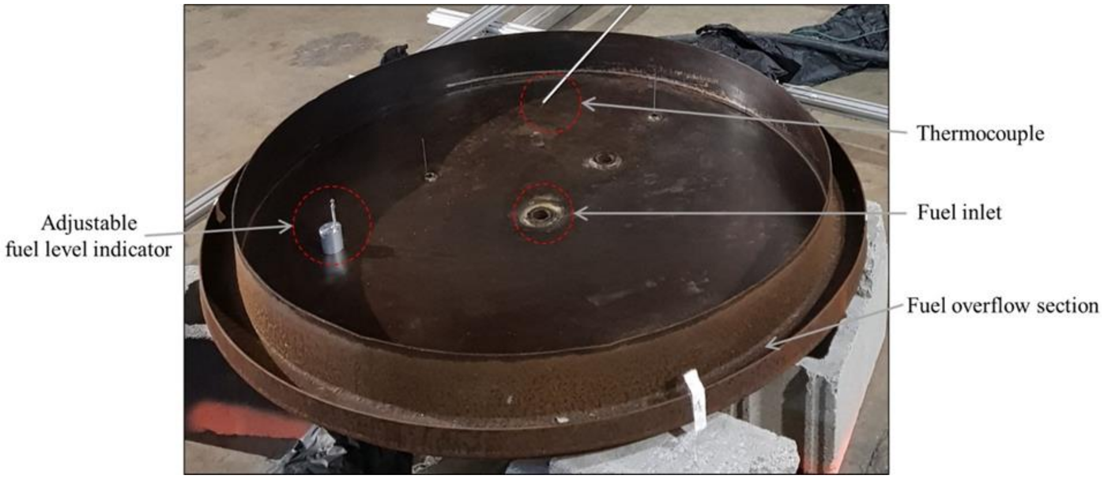


Fig. 1. The 1 m diameter, water-cooled, round steel burner with fuel level indicator and fuel overflow section. The S type thermocouple used to measure the gas phase temperature is also shown.

2.2. Measuring Heat flux

The radiative heat flux by the fire emitted to the surroundings was measured using a wide view angle, water-cooled, Gardon type total heat flux gauges with a 1.3 cm diameter face. The gauges were positioned as shown in Fig. 2. Radial heat flux gauges oriented upward were aligned with the burner rim to measure the heat flux towards the floor. Vertical heat flux gauges were used to measure heat flux to the surroundings.

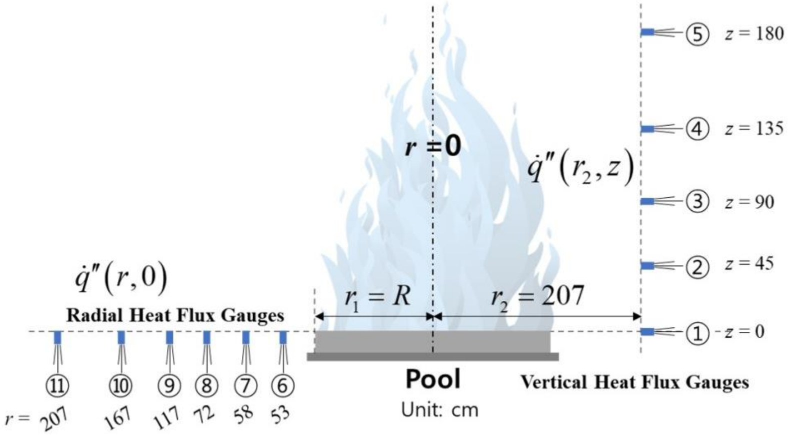


Fig. 2. A schematic diagram of the heat flux gauge set-up. All units in the figure are in cm.

2.3. Measuring Temperature

The local temperature was measured using a Type S (Pt, 10 % Rh/Pt), bare-wire, fine diameter thermocouple. The thermocouple was inserted into a (2-hole) 3 mm outer diameter ceramic tube with about 1 cm of the thermocouple wire including its bead, extending beyond the end of the ceramic tube. Selection of the diameter of a fine wire thermocouple must consider trade-offs between the durability of the instrument and measurement needs. The finer the wire, the smaller the radiative exchange with the environment and the faster the measurement time response, but the more difficult it is to configure. In this study, a 50 µm diameter S-type thermocouple was employed with an approximately spherical bead as observed using an optical microscope. The measured signal was acquired at a rate of 60 Hz for 120 s using a data acquisition module (SCXI-1600, National Instruments Inc.), which represents about 170 flame puffing cycles.

A computer-controlled translation device was used to adjust the position of the thermocouple along a vertical axis aligned with the pool centerline. The vertical rail was aligned with the centerline of the burner and the thermocouple/ceramic tube assembly was attached to the tip of a horizontal rod connected to the moving rail. The connection region between the thermocouple and the rod was well-insulated and covered with aluminum foil.

The energy balance on the thermocouple bead considers convective, radiative, and conductive heat transfer, and can be expressed as:

$$\dot{Q}_{conv} + \dot{Q}_{rad} = \rho_b \cdot c_{p,b} \cdot V_b \frac{dT_b}{dt} \tag{1}$$

where $\dot{Q}$ is the net rate of heat transfer. ρ, $c_{p,b}$, and V_b are the density, specific heat and volume of the bead, respectively. In addition, if the response time of the thermocouple is much larger than the fire fluctuation frequency, then thermal inertia effects can impact the measurement variance, although there is little influence on the mean [4]. The thermal inertia is related to the thermocouple time constant (τ), and the energy balance becomes:

$$T_g(t) = T_b(t) + \tau \frac{dT_b(t)}{dt} + \frac{\varepsilon\sigma}{h}\left(T_b^4(t) - T_{surr}^4\right) \tag{2}$$

$$\tau = \frac{m_b c_{p,b}}{hA_b} \tag{3}$$

where T_b is the bead temperature, T_g is the gas temperature, T_{surr} is the effective temperature of the surroundings, A_b is the surface area of the bead, σ is the Stefan-Boltzmann constant ($5.67 \cdot 10^{-8}$ W/m^2/K^4), ε is the thermocouple emissivity. Here, the flame is taken as essentially optically thin based on estimates using the radiation subroutine in Ref. [6]. The convective heat transfer coefficient of gas flow near the bead is defined as h = Nu $\cdot\ \lambda_g/d_b$, where λ_g is the thermal conductivity of gas, d_b is the thermocouple bead diameter. In Eq. (2), the second and third terms on the right side represent the thermal inertia correction and radiation correction, respectively. The Nusselt number is empirically associated with the Reynolds and Prandtl numbers. Solving the thermal inertia correction term, the time derivative of bead temperature was calculated using a second-order polynomial fit of three consecutive data points of the temperature time series with a curve fit window size of 33.3 ms.

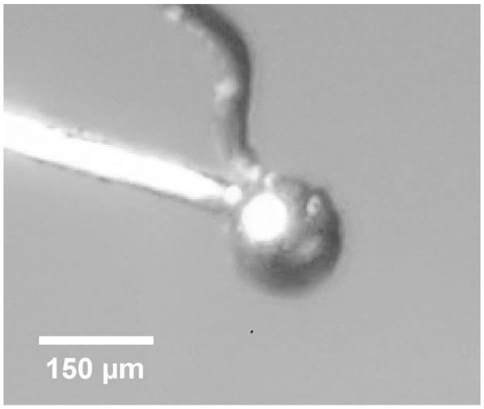


Fig. 3. Magnified image of thermocouple bead.

Fig. 3 shows an image of the thermocouple bead, which was approximately spherical with an eccentricity of about 0.97. The bead diameter was measured using Image-J image processing software from a photo taken with an optical microscope. The uncertainty of the bead diameter was multiplied by the image resolution (2.7 µm/pixel) and the number of pixels needed to determine the edge of the bead. The measured bead diameter was 153.3 µm ± 7.7 µm, which was approximately three times the wire diameter. The time constant for heat transfer to a sphere [14] can be written as:

$$\tau = \frac{\rho_b c_{p,b} {d_b}^2}{6\mathrm{Nu}\lambda_g} \qquad (4)$$

Following Shaddix [15], the Nusselt number for a sphere is calculated using the Ranz-Marshall model:

$$\mathrm{Nu} = 2.0 + 0.6\ \mathrm{Re}^{1/2}\ \mathrm{Pr}^{1/3}; \qquad 0 < \mathrm{Re} < 200 \qquad (5)$$

where Re is the Reynolds number of the bead and Pr is the Prandtl number. The temperature-dependent gas properties for Re and Pr, are taken as those of air [16], and the temperature-dependent emissivity and thermophysical properties of platinum are taken from Refs. [15, 17]. The average ambient temperature during the experiments was 298 K ± 5 K, which was taken as the surrounding temperature, T_{surr}, in Eq. (2). A FDS simulation of the fire was conducted to validate the temperature correction method used to solve Eq. (2) and to obtain the gas velocity distribution above the burner to better represent Re in Eq. (5). The FDS input code was based on the FDS Validation Guide's [6] input file for the 1 m methanol pool fire case. Details are explained in Ref. [18]. The average difference in the mean gas temperature along the centerline between FDS and the experimental results was 4 %. FDS yielded Re ranging from 1 to 24 along the centerline. In Eq. (2), the radiation correction and thermal inertia correction terms mainly affect the mean and variance values, respectively, in agreement with Refs. [4, 15]. For example, correction of the mean temperature due to radiative loss along the centerline was 1 % on average, varying from near zero at the top of the fire plume to 1.7 % at the hottest fire locations. The thermal inertia correction term has a negligible influence on the mean gas temperature, but does amplify the value of its instantaneous extremes, which affects the local standard deviation. The average contribution of the thermal inertia correction term for locations along the centerline represents 54 % of the standard deviation of the gas temperature. In contrast, the radiative loss term has little influence. For these reasons, the uncertainties of the mean and standard deviation of the gas temperature were separately analyzed. The uncertainties of each term of the gas temperature in Eq. (2) were determined based on Ref. [19]. The calibration error of a Type S thermocouple is 0.25 % in 273 K $< T_b <$ 1733 K [20]. The measurement uncertainty of the data acquisition (DAQ) system was approximately 0.60 % for the application range of the thermocouple [21].

3. Results and Discussion

The shape of the fire dramatically changed during its pulsing cycle. The fire was blue with no indication of the presence of soot. Fig. 4 shows four images of the methanol pool fire during different phases of its puffing cycle. Repeating puffing cycles occurred in which orderly curved flame sheets anchored at the burner rim were connected to the central fire plume, rolled towards the fire centerline, and necked-in to form a narrow and long visible fire plume. The flame height was recorded with 30 Hz video. Analysis of the video record showed that the average flame height and its standard deviation was 1.10 m ± 0.22 m and the primary pulsation frequency was 1.37 Hz ± 0.03 Hz.

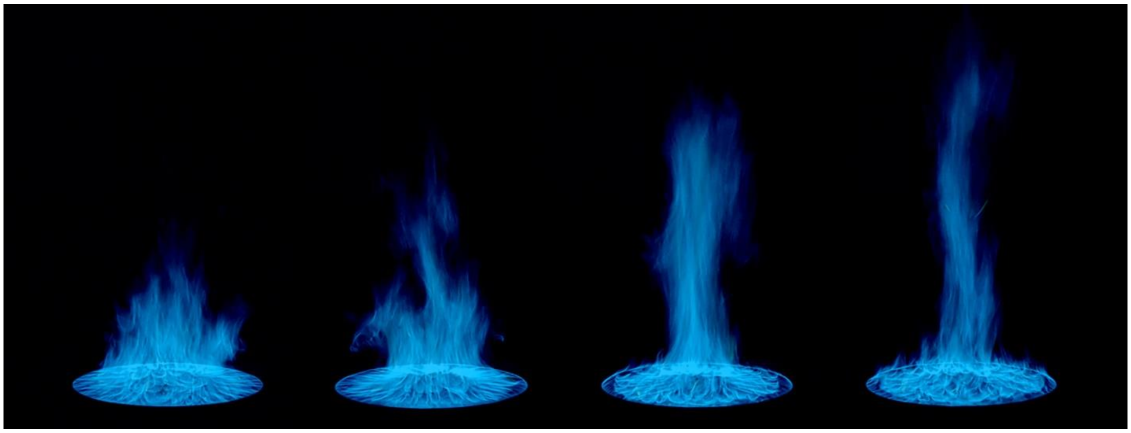


Fig. 4. Instantaneous digital images 132 ms apart in the pulsing 1 m diameter methanol pool fire.

3.1. Mass Burning Rate

With a steady liquid level in the fuel pool, the mass burning rate was measured by monitoring the mass loss in the 20 L methanol reservoir feeding the liquid pool, using a calibrated load cell. Fig. 5 shows the time-varying fuel mass in the reservoir during Test 3. When the fuel was low in the reservoir, it needed to be replenished. The periods when the reservoir was refilled are indicated by the white (unshaded) regions in Fig. 5. During these periods, the fuel was still fed to the burning pool and the fuel level in the pool was maintained constant as verified by a video camera focused on the relative level of the fuel compared to the fuel level indicator (see Fig 1). The burning rate is estimated during the gray regions in the figure, that is, after an initial warm-up and avoiding periods when fuel was added to the reservoir. The total mass loss rate for each period is noted (by the numbers in the gray regions) by considering the ratio of the mass loss to the duration of the period. The time-weighted mean mass burning rate during the three tests was 12.8 g/s ± 0.9 g/s, where the uncertainty here is reported as the combined expanded uncertainty, representing a 95 % confidence interval (a coverage factor of two).

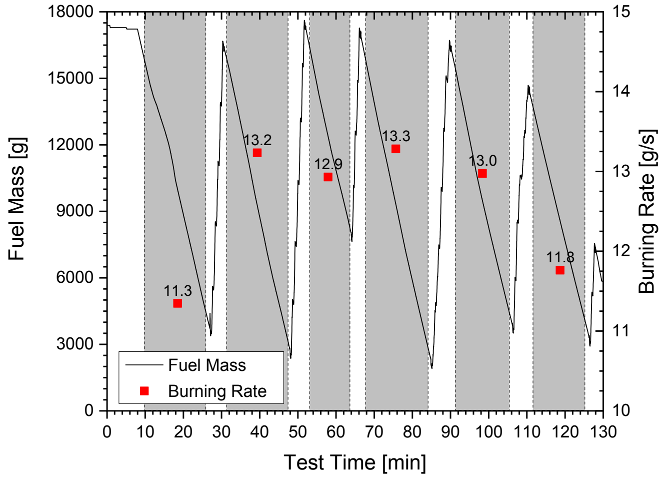


Fig. 5. Mass of fuel reservoir and average fuel burning rate during Test 3. The unshaded regions after 10 min represent times when the reservoir was being refilled with methanol.

3.2. Heat Release Rate

The heat release rate was measured using oxygen consumption calorimetry and compared with the ideal heat release rate ($\dot{Q}$) calculated from the mass burning rate, i.e., $\dot{m}\Delta H_{c,}$ where ΔH_c is the net heat of combustion of methanol equal to 19.9 kJ/g [16]. The heat release rate from calorimetry was averaged for the three tests once the fire reached steady-state burning.

The measured mass burning rate, the ideal heat release rate, and heat release rate measured via the oxygen consumption calorimetry are presented in Table 1. As expected, the ideal heat release rate agrees well with the measured calorimetric heat release rate since the combustion efficiency is expected to be nearly 1. The heat release rate measured by calorimetry was 256 kW ± 45 kW, where the combined expanded uncertainty was based on repeat measurements, the results described in Ref. [13], and additional natural gas calibrations (at a measured heat release rate of about 250 kW).

Table 1. Measured mass burning rate in the 1 m methanol pool fire, the ideal heat release rate determined from the measured mass burning rate, and the heat release rate determined using calorimetry. The uncertainty is expressed as the combined expanded uncertainty with a coverage factor of two, representing a 95 % confidence interval.

Mass burning rate $\dot{m}$ [g/s]	Ideal Heat Release Rate $\dot{Q}$ [kW]	Heat Release Rate from calorimetry $\dot{Q}_a$ [kW]
12.8 ± 0.9	254 ± 19	256 ± 45

3.3. Heat Flux Distribution

Fig. 6 shows the mean radial radiative heat flux as a function of the radial distance from the burner centerline. As expected, the radiative heat flux rapidly decreases with distance from the centerline. The maximum measured radial heat flux was 5.1 kW/m^2 ± 1.0 kW/m^2. The heat flux consistently decreased in a manner proportional to 1/r^2. Fig. 7 shows the mean vertical radiative heat flux as a function of the axial distance above the burner. There was little change in radiative heat flux in the axial direction. The radial heat flux has a maximum value of 1.0 kW/m^2 ± 0.1 kW/m^2 at 0.9 m above the burner. Fig. 6 also shows the results from Ref. [11], which are in agreement with the current measurements within experimental uncertainty.

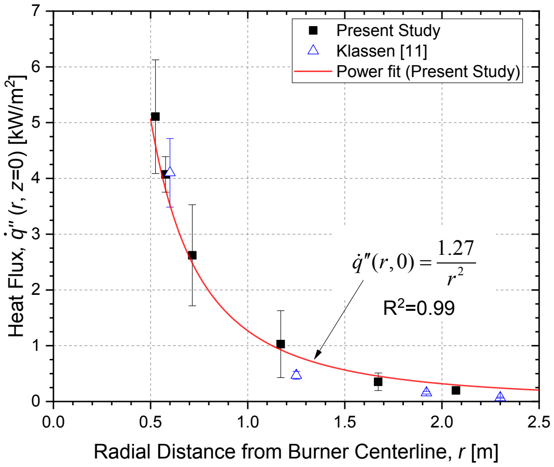


Fig. 6. Mean and standard deviation of the radial radiative heat flux as a function of the radial distance from the burner centerline at the plane defined by the burner rim (z = 0).

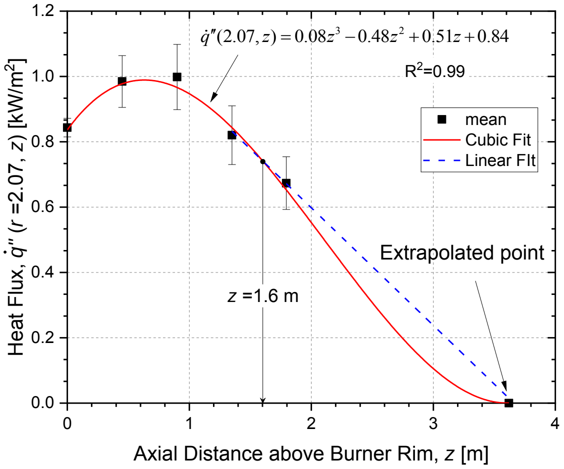


Fig. 7. Mean and standard deviation of the vertical radiative heat flux as a function of the axial distance above the burner for gauges facing the pool fire.

The fraction of energy radiated from the fire (χ_{rad}) was calculated as shown in Eqs. (6) and (7) , considering the overall enthalpy balance explained in Ref. [22], where its value is equal to the ratio of the total radiative emission from the fire $\dot{Q}_{rad}$ normalized by the idealized fire heat release rate ($\dot{Q}$). The radiative fraction can be broken into the sum of the radiative heat transfer to the surroundings (χ_r) and onto the fuel surface (χ_{sr}) such that:

$$\chi_{rad} = \chi_r + \chi_{sr} = \dot{Q}_{rad}/\dot{Q} \tag{6}$$

$$\chi_r = \dot{Q}_r/\dot{Q} \text{ and } \chi_{sr} = \dot{Q}_{sr}/\dot{Q} \tag{7}$$

where $\dot{Q}_r$ is the radiative energy emitted by the fire to the surroundings except to the fuel surface and $\dot{Q}_{sr}$ is the radiative heat feedback to the fuel surface. Assuming symmetry, integrating the measured local radiative heat flux in the *r* and *z* directions (see Fig. 2) yields the total energy radiated by the fire, $\dot{Q}_{rad}$, considering the flux through a cylindrical control surface about the pool fire:

$$\dot{Q}_{rad} = \dot{Q}_r + \dot{Q}_{sr} = \left(2\pi\int_{r_1}^{r_2}\dot{q}''(r,0)\cdot rdr + 2\pi r_2\int_0^{z_2}\dot{q}''(r_2,z)dz\right) + \pi r_1^2\bar{\dot{q}}_{sr}'' \tag{8}$$

where r_1 and r_2 are 0.5 m and 2.07 m, z_2 is 3.62 m, and $\bar{\dot{q}}_{sr}''$ is the average radiative heat flux incident on the fuel surface. In the energy balance for a steadily burning pool fire following Ref. [22], the total heat feedback ($\dot{Q}_s$) to the fuel surface is broken into radiative and convective components ($\dot{Q}_s = \dot{Q}_{sr} + \dot{Q}_{sc}$). Normalizing this by $\dot{Q}$, $\chi_s = \chi_{sr} + \chi_{sc}$. Kim *et al.* [22] measured the distribution of local heat flux incident on the fuel surface in a 30 cm methanol pool fire. The fractional total heat feedback (χ_s) was 0.082 ± 24 % with about 67 % attributed to radiation, that is, χ_{sr}= 0.055 ± 21 %. χ_s in the 1 m pool fire is assumed to be the same as in the 30 cm pool fire. Convective heat transfer to the fuel surface ($\dot{Q}_{sc}$) was calculated using the thin film theory following [23]. As a result, χ_{sr} was 0.065 ± 31 % and χ_{sr}/χ_s was 0.80, which is about 20 % larger compared than in 30 cm pool fire. The fitting function seen in Figs. 6 and 7 was used to integrate the heat flux in the radial and vertical directions. The zero-heat flux position (z_2 = 3.62 m) was extrapolated from the values of the highest two locations in Fig. 7. In previous studies [11, 22], the heat flux peaked at a vertical position equal to approximately one-half the characteristic flame height and decreased almost linearly above the visible flame tip regardless of pool diameter and fuel type, until it reached zero. The vertical radiative heat flux (the second term in Eq. (8)) was integrated using the cubic function from 0 to z_1 (1.6 m) and either the cubic function or a line in the region from z_1 to z_2. The energy difference associated with the fitting functions was treated as uncertainty.

The results show that $\dot{Q}_{rad}$ was 56 kW ± 11 % and χ_{rad} was 0.22 ± 16 %. The radiative fraction of the total heat release rate emitted to the surroundings in previous studies for methanol pool fires is listed in Table 2. The radiative fraction reported here agrees with the value in Ref. [11] within expanded uncertainty. The radiative fraction of the 1 m pool fire was similar to its value in the 30 cm fire, and agreed with the result in Ref. [22] which suggested that the radiative fraction was fairly constant as a function of pool size for diameters less than 2 m.

Table 2. Comparison of the radiative fraction in steadily burning 30 cm and 100 cm methanol pool fires. The combined expanded uncertainty is also shown, representing a 95 % confidence interval.

Research	Pool diameter	χ_{rad}
Present study	100 cm	0.22 ± 16 %
Klassen and Gore [11]	100 cm	0.19[a,b]
Kim *et al.* [22]	30 cm	0.24 ± 25 %
Hamins *et al.* [24]	30 cm	0.22 ± 10 %

[a] $\bar{\dot{q}}_{sr}''$ in Eq. (8) was assumed equal to the heat flux measured next to the burner ($\dot{q}''$(51 cm, 0) = 4.1 kW/m^2), which yields χ_{sr} = 0.01, which is smaller than expected [22]. χ_{rad}, therefore, was recalculated with χ_{sr} = 0.055, yielding χ_{rad} = 0.19.

[b] Recalculated χ_{rad}, using ΔH_c = 19.918 kJ/g [16], not 22.37 kJ/g, assuming gaseous water as a product of combustion.

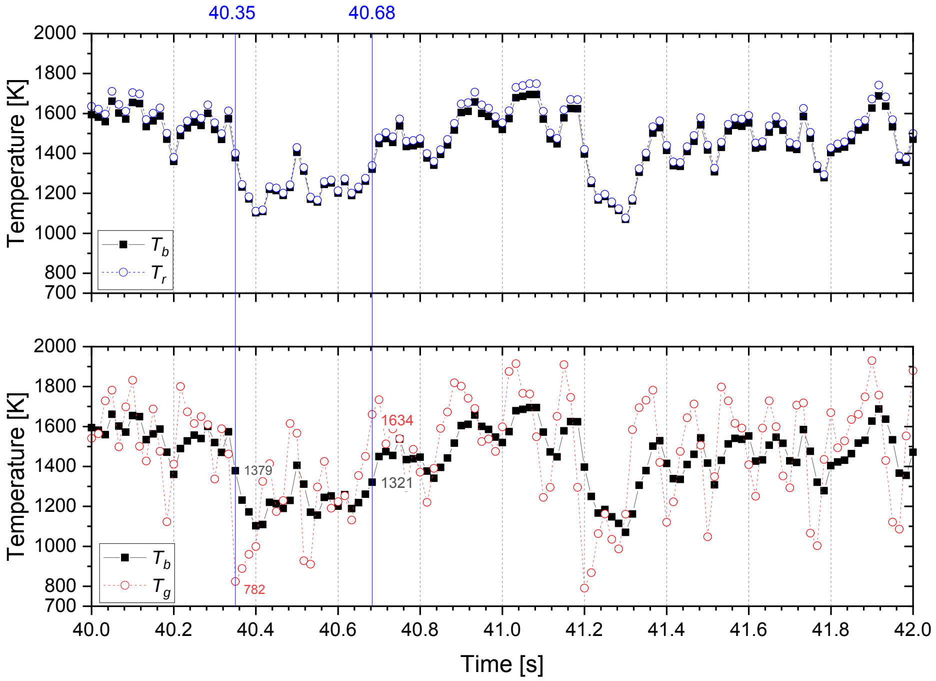
3.4. Temperature Distribution

Fig. 8 shows the measured time series of uncorrected bead temperature (T_b), the radiation corrected temperature (T_r) considering only the radiation correction term (not the thermal inertia) in Eq. (2), and the (radiation and inertia) corrected gas temperature (T_g). There is no time-delay between the bead temperature and the radiation corrected temperature. The radiative correction became larger as the bead temperature increased with the maximum correction equal to 55 K, when T_b =1694 K. The minimum correction was 7 K, when T_b =1070 K in Fig. 8. The corrected gas temperature was 617 K lower than the bead temperature at 40.35 s, whereas it was 313 K higher than the bead temperature at 40.68 s. The mean time constant was calculated as 57 ms ± 3 ms. As the Nusselt number increases with bead temperature, the time constant decreases, as indicated by Eq. (4).

Fig. 9 shows the measured mean and standard deviation of the bead temperature, corrected gas temperature and time constant as a function of distance above the burner along the centerline of the fire in Test 3. As expected, the mean gas temperatures were very similar to the mean bead temperature for all positions. On average, the combined expanded uncertainty of the mean gas temperature was 8 %, considering all 46 temperature measurement locations. On average, the combined expanded uncertainty of the standard deviation of gas temperature as 26 %.

The mean and standard deviation of the gas temperature as a function of distance above the burner along the centerline are shown in Fig. 10. The maximum value of the mean temperature was about 1371 K, which occurred at 0.3 m above the burner rim. The gradient near the fuel

294 surface in Fig. 10 is steep. At 0.05 m above the burner, the gas temperature was about 1144 K ±
295 424 K. The temperature at two locations on the fuel surface was measured to be at the boiling
296 point of methanol, 338 K, yielding a temperature gradient near the fuel surface of about
297 161 K/cm ± 85 K/cm.



298

299 Fig. 8. Instantaneous temperature at (z, r) = (30 cm, 0 cm) in Test 3; T_b is the bead temperature,
300 T_r is the corrected temperature considering only radiative loss, and T_g is the gas temperature
301 corrected for radiative loss and thermal inertia.

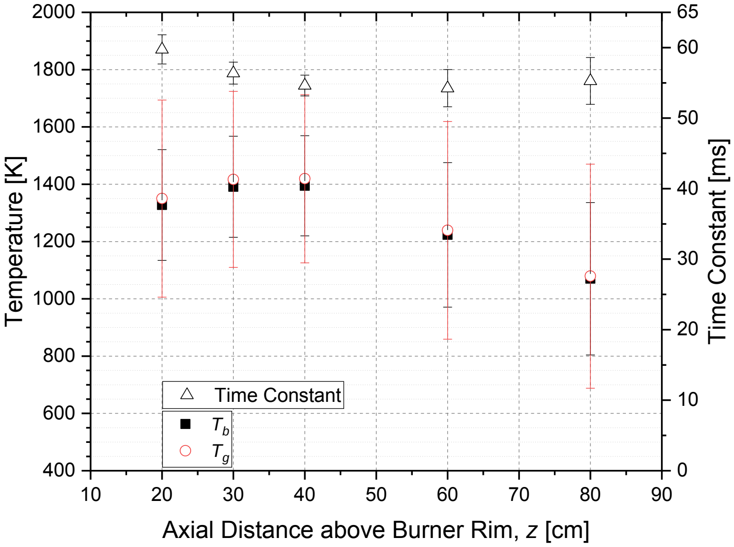


Fig. 9. Mean and standard deviation of the measured bead temperature profile, and calculated gas temperature and thermocouple time constant as a function of axial distance above the burner rim in Test 3.

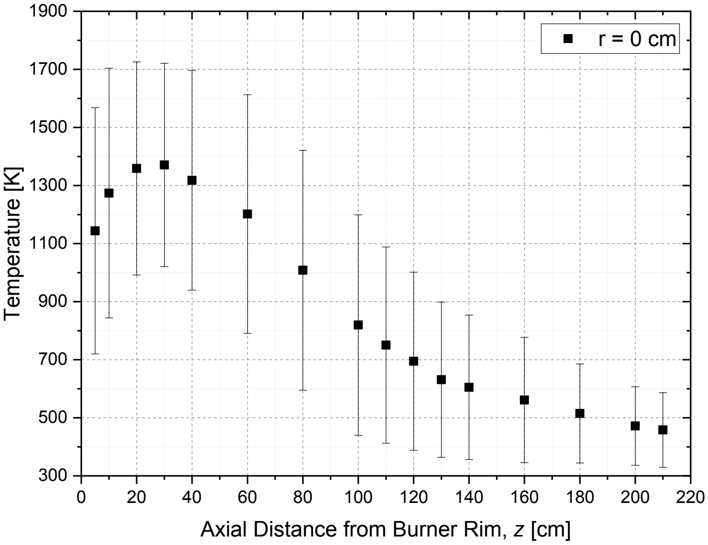


Fig. 10. Mean and standard deviation of the gas temperature profile as a function of axial distance above the burner rim along the centerline of the fire.

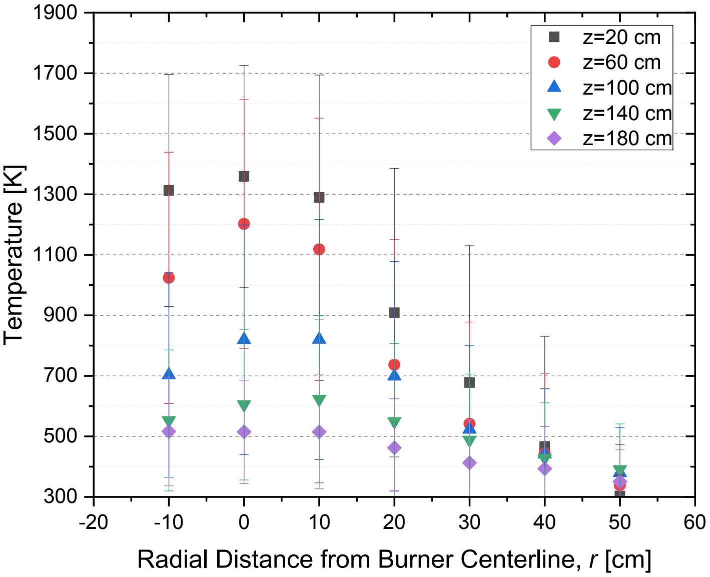


Fig. 11. Mean and standard deviation of the gas temperature profiles as a function of radial distance from the burner centerline at various heights above the burner rim.

Fig. 11 shows the mean and standard deviation of the gas temperature profile in the radial direction for various axial distances above the burner rim (20 cm ≤ z ≤ 180 cm). The maximum temperature occurs near the centerline for each elevation. The gradient diminished with distance from the fuel surface. A complete discussion of the uncertainty analysis for the temperature and other results is given in Ref. [18].

Sung, Kunhyuk; Chen, Jian; Bundy, Matthew; Hamins, Anthony. "The Characteristics of a 1 m Methanol Pool Fire." Paper presented at 13 International Symposium on Fire Safety Science (IAFSS 2020), Waterloo, CA. April 26, 2020 - May 01, 2020.

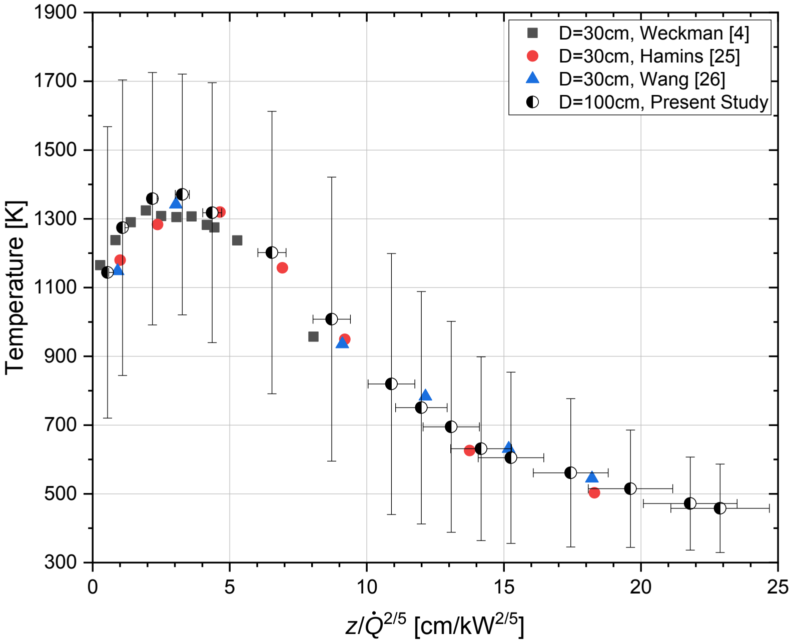


Fig. 12. Mean and standard deviation of the axial temperature profiles as a function of distance above the burner rim normalized by $\dot{Q}^{2/5}$ and compared to previous results in 30 cm methanol pool fires.

Fig. 12 shows the mean and standard deviation of the temperature profile as a function of scaled axial distance. The results are compared to previous measurements in 30 cm diameter methanol pool fires from Refs. [4, 25, 26]. Axial distance above the burner is normalized by $\dot{Q}^{2/5}$ following Baum and McCaffrey [27]. Weckman and Strong [4] measured temperature in a 30.5 cm diameter methanol pool fire with a lip height of 1 cm using a 50 µm wire diameter, bare bead, Type S (Pt, 10% Rh/Pt), thermocouple similar to the thermocouples used in this study. The measurements from Ref. [25] are also shown, where temperature was measured using a 75 µm wire diameter, bare bead, Type S thermocouple in a steadily burning 30.1 cm diameter methanol pool fire with a 0.6 cm lip. The radiation corrected thermocouple measurements in Wang *et al.* [26] are also shown, using a 50 µm wire diameter, bare bead, Type S thermocouple in a steadily burning 30.1 cm diameter methanol pool fire with a 1 cm lip height. A comparison of the results in Fig. 12 shows that the 1 m and 30 cm pool temperatures are similar when the axial distance above the burner is normalized by $\dot{Q}^{2/5}$.

4. Summary and Conclusions

A series of measurements for temperature, burning rate and heat release rate were conducted to characterize a 1 m diameter, well-ventilated methanol pool fire steadily burning in a quiescent environment. The measured heat release rate determined by oxygen consumption calorimetry was 256 kW ± 45 kW, which was consistent with the heat release rate calculated from the fuel mass burning rate measurements. The gas-phase thermocouple temperature measurements were corrected considering radiative loss and thermal inertia effects. Instantaneous temperatures as large as 1800 K were measured in the fire. The maximum value of the time-averaged gas temperature was measured as about 1371 K, which occurred about 0.3 m above the burner. As expected, the corrected profile of mean axial temperature was shown to be similar to previous results for methanol pool fires when scaled by $\dot{Q}^{2/5}$. The heat flux was measured in the radial and vertical directions, and the radiative fraction was estimated as 0.22 ± 16 %, which corresponded to previous methanol pool fire results in 1 m and 0.3 m diameter pools. The present results help provide an understanding of the structure and character of the 1 m diameter methanol pool fire and provide data useful for the evaluation of fire models.

Acknowledgements

The authors are grateful to Marco Fernandez of NIST for assistance with the measurements.

References

[1] K Akita, T Yumoto, Heat transfer in small pools and rates of burning of liquid methanol, Proceedings of the Combustion Institute 10 (1965) 943-948.

[2] A Hamins, S J Fischer, T Kashiwagi, M E Klassen, J P Gore, Heat feedback to the fuel surface in pool fires, Combustion Science and Technology 97 (1994) 37-62.

[3] S Hostikka, K B Mcgrattan, A Hamins, Numerical modeling of pool fires using LES and finite volume method for radiation, Fire Safety Science 7 (2003) 383-394.

[4] E J Weckman, A B Strong, Experimental investigation of the turbulence structure of medium-scale methanol pool fires, Combustion and Flame 105 (1996) 245-266.

[5] A Yilmaz, Radiation transport measurements in methanol pool fires with Fourier transform infrared spectroscopy, NIST Grant/Contractor Report GCR 09-922, January 2009.

[6] K McGrattan, R McDermott, M Vanella, S Hostikka, J Floyd, C Weinschenk, K Overholt, Fire dynamics simulator user's guide, NIST special publication 1019, National Institute of Standards and Technology, October 2019.

[7] G Maragkos, T Beji, B Merci, Towards predictive simulations of gaseous pool fires, Proceedings of the Combustion Institute 37 (2019) 3927-3934.

[8] Z Chen, J Wen, B Xu, S Dembele, Large eddy simulation of a medium-scale methanol pool fire using the extended eddy dissipation concept, International Journal of Heat and Mass Transfer 70 (2014) 389-408.

[9] S R Tieszen, T J O'Hern, R W Schefer, E J Weckman, T K Blanchat, Experimental study of the flow field in and around a one meter diameter methane fire, Combustion and Flame 129 (2002) 378-391.

[10] S R Tieszen, T J O'Hern, E J Weckman, R W Schefer, Experimental study of the effect of fuel mass flux on a 1-m-diameter methane fire and comparison with a hydrogen fire, Combustion and Flame 139 (2004) 126-141.

[11] M Klassen, J Gore, Structure and radiation properties of pool fires, NIST-GCR-94-651, National Institute of Standards and Technology, Gaithersburg, MD, June 1994.

[12] R D Peacock, W Jones, P Reneke, G Forney, CFAST–consolidated model of fire growth and smoke transport (version 6) user's guide, NIST special publication 1041r1 (2013)

[13] R A Bryant, M F Bundy, The NIST 20 MW calorimetry measurement system for large-fire research, NIST Technical Note 2077, National Institute of Standards and Technology, Gaithersburg, MD, 2019.

[14] T L Bergman, F P Incropera, D P DeWitt, A S Lavine, Fundamentals of heat and mass transfer, John Wiley & Sons, 2011.

[15] C R Shaddix, Correcting thermocouple measurements for radiation loss: A critical review, American Society of Mechanical Engineers, New York, Sandia National Labs., Livermore, CA (US), 1999.

[16] Design institute for physical properties (DIPPR 801), American Institute of Chemical Engineers, 2017.

[17] F M Jaeger, E Rosenbohm, The exact formulae for the true and mean specific heats of platinum between 0 ° and 1600 °C, Physica 6 (1939) 1123-1125.

[18] K Sung, J Chen, M Bundy, M Fernandez, A Hamins, The thermal character of a 1 m methanol pool fire, NIST Technical Note 2083, National Institute of Standards and Technology, Gaithersburg, MD, 2020.

[19] B N Taylor, C E Kuyatt, Guidelines for evaluating and expressing the uncertainty of NIST measurement results, NIST Technical Note 1297, National Institute of Standards and Technology, Gaithersburg, MD, 1994.

[20] The temperature handbook, Omega Engineering Inc., 2004, pp. Z-39-40.

[21] SCXI-1600 user manual and specifications, National Instruments Inc., 2004.

[22] S C Kim, K Y Lee, A Hamins, Energy balance in medium-scale methanol, ethanol, and acetone pool fires, Fire Safety Journal 107 (2019) 44-53.

[23] L Orloff, J de Ris, Froude modeling of pool fires, Technical Report FMRC OHON3.BU, RC81-BT-9, Factory Mutual Research Corp., Norwood, MA, 1983.

[24] A Hamins, M Klassen, J Gore, T Kashiwagi, Estimate of flame radiance via a single location measurement in liquid pool fires, Combustion and Flame 86 (1991) 223-228.

[25] A Hamins, A Lock, The structure of a moderate-scale methanol pool fire, NIST Technical Note 1928, National Institute of Standards and Technology, Gaithersburg, MD, 2016.

[26] Z Wang, W C Tam, K Y Lee, A Hamins, Temperature field measurements using thin filament pyrometry in a medium-scale methanol pool fire, NIST Technical Note 2031, National Institute of Standards and Technology, Gaithersburg, MD, 2018.

[27] H R Baum, B McCaffrey, Fire induced flow field-theory and experiment, Fire Safety Science 2 (1989) 129-148.

AccuPIPE: Accurate Heavy Flow Detection in the Data Plane Using Programmable Switches

Yang Guo, *NIST*
Franklin Liu, *UIUC*
An Wang, *Case Western Reserve University*
Hang Liu, *The Catholic University of America*

***Abstract*—Identifying heavy flows, i.e., flows with large packet counts during a pre-defined time window, is vital for many network applications. The task of real-time heavy flow detection in data plane is challenging due to high switching speed (100 Gbps), a large number of concurrent flows (millions of concurrent flows), and small memory footprint requirement. In this paper, we dissect the key factors that affect the existing detection scheme's accuracy, and propose AccuPipe, a new detection scheme with intelligent flow entry replacement strategies. The simulation results show that the new scheme is able to efficiently utilize all flow entries in the detection pipeline, and detects more than 850 heavy flows (out of top 1,000) using a small amount of memory (1,000 flow entries, roughly equivalently to 18KB memory) with reasonable reporting overhead. This represents a 76% improvement over HashPIPE scheme, which detects on average 484 heavy flows (out of top 1,000) in the same setting. In addition, we investigate the performance of different flow entry replacement strategies, and report their pros and cons.**

I. Introduction

In the Internet, a small set of heavy flows, flows with extremely large number of packets or bytes, often accounts for a disproportionate share of the total traffic. Real-time detection of such heavy flows at small time scales is useful for many applications, e.g., DDoS detection, dynamic traffic routing [14], dynamice flow scheduling [15], etc. Real-time heavy flow detection in the data plane, however, is challenging due to stringent speed, accuracy, and memory requirement.

Heavy flow detection is well studied in streaming algorithms literature [6], where packets are processed as they pass through the measurement point. Packet sampling [5], [8], sketching [7], [9], [17], and counter-based algorithms [11] are representative techniques that tackle the trade-off between measurement accuracy, speed, and momory usage. For instance, sample and hold [8] imporves the sampling accuracy by keeping counters for flows that have been sampled. Count-min sketch [7] hashes on packet headers and increments counters in hash tables. The minimum counter is used to approximate the flow packet counts. FlowRadar [9] improves upon count-min sketch by allowing keys (flow IDs) to be decoded from the hash. More recently, UnivMon [10] and Elastic Sketch [18] develop sketch based techniques where one sketch satisfies multiple tasks' requirements.

HashPipe [16] is the first to investigate the real-time heavy flow detection problem in a switch's data-plane. HashPipe takes advantage of emerging programmable switch's programmability [1]–[3] and designs a measurement pipeline that implements modified space saving [12] algorithm to capture heavy flows. Elastic sketch [18] is another in-data-plane flow measurement scheme that utilizes more memories than HashPipe but provides the flow rate estimations for all flows.

In this paper we strive to improve upon HashPipe to achieve better heavy flow detection accuracy with small memory footprint. We analyze two key factors that affect HashPipe's measurement accuracy, namely *flow entry wastage* in the measurement pipeline and *packet arrival pattern*. Our analysis and simulation results show that the flow entry wastage contributes less than 8% to the accuracy loss, while the packet arrival pattern plays a major role in the measurement inaccuracy due to dramatic time-varying packet arrival pattern at fine time scale.

We propose AccuPipe that integrates frequency-based caching replacement strategies into heavy flow measurement pipeline. Instead of using the accumulated packet count as the indicator of potential heavy flows as in HashPipe, *AccuPipe utilizes flow entries inside the measurement pipeline to opportunistically capture short-term packet bursts and reports them to the measurement server.* At the end of each measurement cycle, the measurement server aggregates the reported packet counts to identify the top heavy flows. Our results show that the proposed scheme improves the identification accuracy from 48% to more than 85% with reasonable reporting overhead.

The rest of the paper is organized as follows. Section II introduces HashPipe and analyzes key factors that affect HashPipe's measurement accuracy. Section III describes AccuPipe and flow entry replacement strategies. Section IV presents the performance evaluation results. Section V concludes the paper.

II. HashPipe and Its Accuracy Analysis

HashPipe uses the measurement pipeline that implements modified space saving [12] algorithm to capture heavy flows. The measurement pipeline is defined using the programming language, e.g., P4 [2] [1] and runs in a programmable switch's data-plane [1]. The pipeline consists of multiple stages, with each stage working as a hash table. The entries in the hash

[1] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Guo, Yang; Liu, Franklin; Wang, An; Liu, Hang. "AccuPIPE: Accurate Heavy Flow Detection in the Data Plane Using Programmable Switches." Paper presented at IEEE/IFIP Network Operations and Management, Budapest, HU. April 20, 2020 - April 24, 2020.

table are called *flow entries*. In HashPipe, a flow entry contains two fields, key field and val field. The key field holds the flow id, e.g., five tuple {`srcIP, dstIP, srcPORT, dstPORT, Protocol`} as defined by the measurement. The val field holds the accumulated packet count of the corresponding flow. Programmable switches use *packet metadata* to communicate the results of packet processing between different stages. The metadata traverses the pipeline together with the packet. In HashPipe, the packet metadata also contains two fields, $cKey$ (*carried Key*) and $cVal$ (*carried value*). The first stage hashes on the key of incoming packet ($iKey$), while the rest of stages hash on $cKey$.

The first stage in HashPipe always allows the incoming packet to be inserted into the hash table. Let $h_i(\cdot)$ be the hash function at stage i. The incoming packet is hashed to flow entry l, $l = h_1(iKey)$. If the entry l is empty and not used, we set $key_l = iKey$ and $val_l = 1$. The packet finishes the measurement and exits the pipeline. If key_l matches $iKey$, the value of val_l is increased by one and the packet also exits the pipeline. If $iKey$ does not match key_l, $(cKey, cVal) \quad (key_l, val_l)$ and $(key_l, val_l) \quad (iKey, 1)$. The packet carries the metadata and goes to the next stage.

The rest of the stages keep heavy flows in the pipeline and push out small flows. Upon the arrival of a packet, the stage i hashes on $cKey$ and gets flow entry l, where $l = h_i(cKey)$. If the flow entry l is empty or $key_l = cKey$, the entry is updated as in the first stage and the packet exits the pipeline. If, however, the key_l does not match $cKey$, $cVal$ is compared with val_l. Whichever key with larger value will be kept in the hash table, and the smaller one is saved in the metadata and moved to the next stage. In the end the flow with the smallest *val* is washed out of HashPipe. The detailed algorithm is included in the Section 3.4 in [16].

A. Accuracy analysis of HashPipe

Flow entry wastage, traffic arrival pattern, and duplicates negatively affect HashPipe's accuracy. Consider a N flow-entry, K stage HashPipe pipeline. We define *heavy flows* to be the flows ranked in the top N based on their packet counts during a pre-defined time window. The rest of flows are called *light flows*. Ideally HashPipe should be able to capture all heavy flows, with one flow entry for one heavy flow. However, since flows are hashed randomly to flow entries, some flow entries may not "see" any heavy flows, and are thus wasted. We denote such behavior as *flow entry wastage*. Also, flow entries can be occupied by light flows instead of heavy flows. A light flow can overtake competing heavy flows with favorable *traffic arrival pattern*, as shown later in this section. Finally, a heavy flow may be hosted at multiple flow entries at different stages. Such *duplicates* also negatively affect HashPipe's accuracy. The simulation study in [16] shows that duplicates account for from 5% to 15% towards inaccuracy. In this section we focus on the effects of *flow entry wastage* and *traffic arrival pattern*.

- **Effect of flow entry wastage.** Below we derive a model that estimates S, the average number of wasted flow entries.

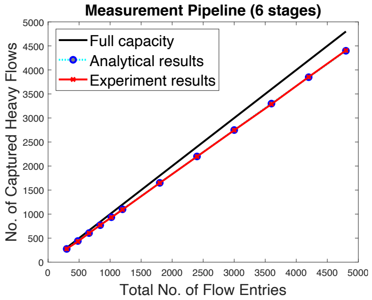


Fig. 1: HashPipe hash collision error analysis.

In the derivation, we ignores the *duplicates* effect and *traffic arrival pattern* effect. We also assume that the first HashPipe stage behaves the same as the later stages. These assumptions favor HashPipe and the analysis offers a lower bound of S.

Given N flow entries and K stages, the average number of flow entries per stage is N/K. Denote by s_i the average number of wasted entries at stage i, and by n_i the average number of incoming heavy flows that are looking to be hosted at stage i. We have $n_1 = N$, and

$$n_i = n_{i-1} - N/K + s_{i-1} \tag{1}$$

for $i = 2, ..., K$. Stage $(i-1)$ has N/K entries with s_{i-1} being wasted. Hence the average number of heavy flows captured at stage $(i-1)$ is $N/K - s_{i-1}$. The number of incoming heavy flows at stage i is thus $n_{i-1} - N/K + s_{i-1}$.

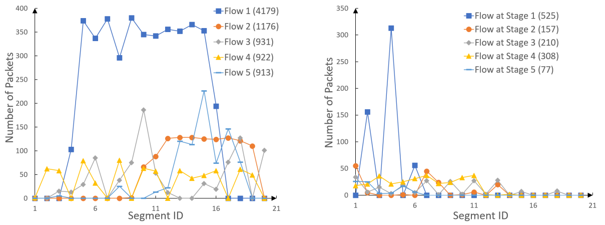
We next derive the formula for s_i. Consider a heavy flow that is randomly hashed to a flow entry at stage i. The probability that a flow entry is not chosen by this heavy flow is $1 - \frac{1}{N/K}$. The probability that a flow entry is not chosen by any n_i heavy flows is $(1 - K/N)^{n_i}$. Hence the average number of wasted entries at stage i is:

$$s_i = (N/K) \cdot (1 - K/N)^{n_i} \tag{2}$$

for $i = 1, 2, ..., K$. Starting with $n_1 = N$, $(\{s_i\}, \{n_i\})$ can be computed iteratively using Eqn.(1) and (2).

Denote by C the average number of captured heavy flows by HashPipe, i.e., $C = N - \sum_{i=1}^{K} s_i$. Figure 1 depicts the average number of captured heavy flows with varying number of total flow entries, The number of stages is set to be six, shown to be optimal in [16]. Varying the number of stages does not change our conclusions. We also conduct the simulations to verify the analytic model. We use a two-second CAIDA trace [4] with one million packets and over 100k flows to drive the simulation. Five-tuple {`srcIP, dstIP, srcPORT, dstPORT, Protocol`} is used as the flow id through out the paper. The simulation results are consistent with the analysis. The HashPipe's measurement pipeline wastes no more than 8% of the flow entries. Using the same trace with 1,000 flow entries, HashPipe only captures about 484 heavy flows, or 48% of top 1,000 heavy-flows. This raises the question if *traffic arrival pattern* plays a more important role than *flow entry wastage* and *duplicates*.

- **Effect of traffic arrival pattern.** We next examine the impact of traffic arrival pattern to the HashPipe accuracy.



(a) Top five heavy flows missed by HashPipe

(b) Flows occupy the flow entry at different stages in HashPipe

Fig. 2: Effect of traffic arrival pattern to the HashPipe accuracy.

To observe the traffic arrival pattern, the two-second trace is divided into 20 segments (0.1 second per segment). Fig. 2a plots the number of packet arrvials per segment for the top five heavy flows missed by HashPipe. The flows are ranked from one to five in the decreasing order of their packet counts, which is included in the parenthesis. Furthermore, Fig. 2b depicts the flows that occupy the flow entries that Flow 1 is hashed to at five pipeline statges (we ignore HashPipe's first stage since it is used for staging the measurement). Flow 1, with the packet count of 4,179, is a much heavier flow than any of the five flows hosted in HashPipe. Flow 1, however, does not arrive until 4th segment, when the other flows already have established a healthy packet count in the pipeline. Flow 1 fails to accumulate a big packet count in the initial stage of HashPipe and loses the competition with smaller flows. We also examine other heavy flows that are not captured by the HashPipe. The results show that the majority of the missed heavy flows are caused by the traffic arrival pattern instead of flow entry wastage and duplicates. It has long been observed that the network traffic is very bursty due to underlying protocol such as TCP, and packet burst varies dramatically at small time scale. A large packet count is not a robust indicator that a flow is heavy. The design of HashPipe favors flows with early burst packet arrival, not necessarily heavy flows.

III. Design of AccuPipe

In this section, we introduce *AccuPipe (Accurate HashPipe)* to accurately captures heavy flows. Unlike HashPipe where packet count is used to predict the heavy flows, AccuPipe uses the measurement pipeline as a cache to opportunistically captures ongoing heavy flows. When an ongoing flow is deemed not heavy anymore, its packet count is evicted from the measurement pipeline and reported to the server. The emptied flow entry continues to capture next heavy flow burst. At the end of a measurement window, the server aggregates collected packet counts and obtains top heavy flows.

In Caching technology, contents are stored in temporary storage, so-called cache, to reduce the service delay or the workload on servers [13]. To maximize the caching hit ratio, various cache replacement strategies have been developed, e.g., recency-based strategies (e.g. LRU), frequency-based strategies (e.g. LFU), function-based strategies, randomized strategies, etc. AccuPipe use the measurement pipeline as a cache to capture the heavy flow packet burst. A set of flows being hashed to the same flow entry compete for its occupancy; replacement strategies are designed to store heavy flows with frequent packet arrivals. Hence frequency-based caching strategies are most suitable for our design. Below we describe several replacement strategies used in AccuPipe.

•**Aging based strategy**: An *age* field is added to flow entries. Upon the arrival of a packet, if $iKey$ matches flow entry key, the packet count is increased by one and the *age* field is reset to be zero, regardless of the current value. The packet then exits the pipeline. If the packet $iKey$ does not match flow entry's key, the *age* field is increased by one. The *age* is then compared to a pre-set threshold α. If the *age* is less than α, the packet continues to traverse the pipeline. On the other hand, if the *age* is greater than $alpha$, the current flow in the flow entry is deemed to be infrequent and evicted. Its packet count is reported to the server. The newly arrived packet is cached into the flow entry and exits the pipeline.

•**Frequency based strategy**: Frequency based strategy was first proposed in [18]. In frequency based strategy, a new field called *total_packet_count* is added to each flow entry. The *total_packet_count* is initialized to be zero, and increased by one whenever a packet is hashed to this flow entry. The ratio of *total_packet_count / packet_count* is computed and compared to the preset threshold λ. The rest of the process is the same as in the Aging based strategy.

•**Segment based strategy**: In segment based strategy, AccuPipe works the same as HashPipe. The measurement time window is divided into equal size small measurement sub-windows, or segment. At the end of each segment, the measurement results in the pipeline are reported to the server. The pipeline is re-initialized and continues the measurement for the next segment.

•**Hybrid strategy**: Hybrid strategy is a combined scheme of aging based strategy and frequency based scheme. If the packet count is less than or equal to the pre-set threshold β, frequency based strategy is applied. If the packet count surpasses β, the aging based strategy is applied. As shown in Section IV, hybrid strategy offers best performance.

IV. Evaluations

In this section, we evaluate AccuPipe with different caching strategies. The measurement pipeline consists of six stages and 1,000 flow entries. The pcap trace is from [4] and is pre-processed to extract ten 5-second traces based on packet time-stamp information. Each 5-second trace contains more than 2.5 million packets and over 200K flows. The results presented here are the average over ten traces. Two metrics are examined: *accuracy*, the number of heavy flows that are correctly identified, and *reporting overhead*, the amount of reports that are sent to the server. Note that heavy flows are the flows being ranked in the top 1,000.

• **Performance comparison of Aging, Frequency, and Segmented AccuPipe.** The aging threshold, frequency threshold, and number of segments are varied in evaluating the

Guo, Yang; Liu, Franklin; Wang, An; Liu, Hang. "AccuPIPE: Accurate Heavy Flow Detection in the Data Plane Using Programmable Switches." Paper presented at IEEE/IFIP Network Operations and Management, Budapest, HU. April 20, 2020 - April 24, 2020.

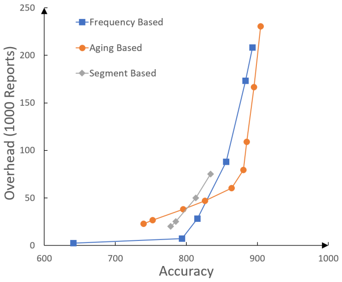


Fig. 3: Performance comparison of Aging, Frequency, and Segmented AccuPipe. Each dot represents the average results over ten traces for one experiment setting. The aging thresholds (α) from the right to the left are 8, 16, 32, 48, 64, 80, 96, 128, and 144. The frequency thresholds (λ) from the right to the left are 2, 4, 8, 16, 32, and 64. In Segmented AccuPipe experiments, the number of segments are 20, 25, 50, and 75, respectively from the left to the right.

accuracy and the reporting overhead of Aging AccuPipe, Frequency AccuPipe, and Segmented AccuPipe. Fig. 3 depicts the reporting overhead vs. accuracy with different thresholds and number of segments, where accuracy represents the number of heavy flows that are correctly identified. As expected, smaller aging and frequency thresholds offer higher accuracy than larger ones as AccuPipe captures smaller traffic bursts and report more frequently. Aging scheme can capture more than 850 heavy flows accurately as long as the aging threshold is smaller than 64, while Frequency scheme can do so if the frequency threshold is no greater than eight. The reporting overhead for Aging and Frequency AccuPipe are 60.1k and 87.9k reports, respectively, which favors aging scheme.

However, as the threshold decreases, frequency based scheme starts to outperform aging scheme. At the frequency threshold of 32, the *accuracy percentage*, the percentage of heavy flows that are correctly identified, reaches 79.3% with only 7.1k reports in Frequency AccuPipe; the similar accuracy percentage (79.5% with the aging threshold of 96) requires 60.1k reports in Aging AccuPipe. In general, aging based scheme performs well in high accuracy percentage ($> 85\%$) region while frequency scheme performs better in medium accuracy region, as shown in Fig. 3.

In Segmented AccuPipe, the packet counts in the pipeline are reported to the switch controller at the end of each segment, and the measurement is restarted for the next segment. We assume that each flow entry requires one report. In general, using more segments in Segmented AccuPipe leads to higher accuracy since the measurement is able to keep track of traffic pattern at the finer time scale. In terms of reporting overhead, Segmented AccuPipe incurs more overheand than Frequency AccuPipe to achieve the same accuracy. Compared to Aging AccuPipe, Segmented AccuPipe incurs more overhead when the accuracy is greater than 800 heavy flows. It, however, starts to outperforms Aging AccuPipe as the accuracy decreases to be smaller than 800 heavy flows. Note that the packet counts in Segmented AccuPipe are periodically reported. Hence these reports form a batch and can potentially be compressed. Compression can drastically reduce the reporting overhead and make Segmented scheme desirable.

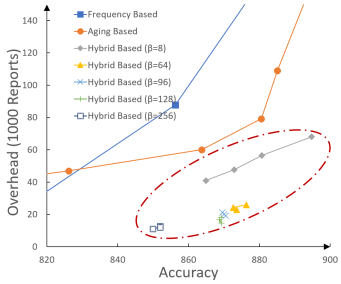


Fig. 4: Performance of Hybrid AccuPipe. The value of α is set at 32 for all cases. The value of β is shown in the figure. The value of λ are set at 32, 64, 96, and 128 from the right to the left. We retain the results of Aging and Frequency AccuPipe for the purpose of comparison.

- **Performance of Hybrid AccuPipe.** Compared to Frequency AccuPipe, Aging AccuPipe is more accurate and requires less reporting overhead in high accuracy region (accuracy percentage $> 85\%$). Frequency AccuPipe, however, is able to bring down the reporting overhead without significantly sacrificing the accuracy in lower accuracy region. Hybrid AccuPipe strives to combines Aging scheme's accuracy with Frequency scheme's reporting efficiency. In the Frequency AccuPipe with the threshold of λ, a flow with the burst arrival of n packets at the beginning can occupy the flow entry for the period of $\lambda \cdot n$ packet arrivals hashed to the same flow entry. A large value of λ could casues the overstay of unheavy flows. In contrast, Aging AccuPipe always examines the most recent α packets. If no packet belonging to the flow entry key arrives during the most recent α packet arrivals, the packet count in the flow entry will be evicted.

To avoid the overstay of a unheavy flows in Frequency AccuPipe, a new threshold β is introduced in Hybrid AccuPipe. Hybrid AccuPipe has three parameters, α, β, and λ. When a flow entry's packet count is less than β, AccuPipe employs Frequency scheme. Otherwise AccuPipe employs Aging scheme. Fig. 4 depicts the accuracy vs. overhead of Hybrid AccuPipe with different values of β. Hybrid AccuPipe consistently outperforms Aging and Frequency AccuPipe. For instance, when $\beta = 64$, Hybrid AccuPipe achieves 87.3% accuracy percentage with 22.9k reports.

V. Conclusions

In this paper, we study the accuracy issue of the real-time heavy flow detection using programmable switch. We investigate two factors that contribute to the measurement inaccuracy, and identify traffic arrival pattern is the main contributor to the inaccuracy in the existing scheme. We propose AccuPipe that employs caching techniques to capture short-term traffic bursts. The simulation results show that the AccuPipe improves the heavy flow detection rate by over 70% over the existing scheme.

References

[1] Barefoot Networks. Barefoot Tofino. https://www.barefootnetworks.com/technology/.

[2] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, and D. Walker. “P4: Programming protocol-independent packet processors.” In ACM SIGCOMM Computer Communication Review, 2014.

[3] P. Bosshart, G. Gibb, H.-S. Kim, G. Varghese, N. McKeown, M. Izzard, F. Mujica, and M. Horowitz. “Forwarding metamorphosis: Fast programmable match-action processing in hardware for SDN.” In ACM SIGCOMM, 2013.

[4] The CAIDA Anonymized Internet Traces 2016 Dataset. http://www.caida.org/data/passive/passive_2016_dataset.xml.

[5] Cisco Networks. Netflow. http://www.cisco.com/c/en/us/products/ios-nx-os-software/ios-netŒow/index.html.

[6] G. Cormode and M. Hadjieleftheriou. “Finding frequent items in data streams.” In VLDB Endowment, 2008.

[7] G. Cormode and S. Muthukrishnan. “An improved data stream summary: The count-min sketch and its applications.” In Journal of Algorithms, 55(1):58–75, 2005.

[8] C. Estan and G. Varghese. “New directions in traffic measurement and accounting.” In ACM Trans. Computer Systems, 21(3), 2003.

[9] Y. Li, R. Miao, C. Kim, and M. Yu. “FlowRadar: A better NetFlow for data centers.” In USENIX NSDI, 2016.

[10] Zaoxing Liu, Antonis Manousis, Gregory Vorsanger, Vyas Sekar, and Vladimir Braverman. “One sketch to rule them all: Rethinking network flow monitoring with univmon.” In Proceedings of the 2016 conference on ACM SIGCOMM 2016 Conference. ACM, 2016.

[11] G. S. Manku and R. Motwani. “Approximate frequency counts over data streams.” In VLDB Endowment, 2002.

[12] A. Metwally, D. Agrawal, and A. El Abbadi. “Efficient computation of frequent and top-k elements in data streams.” In International Conference on Database Theory. Springer, 2005.

[13] Stefan Podlipnig and Laszlo Boszormenyi. “A Survey of Web Cache Replacement Strategies”, ACM Computing Surveys, 2003.

[14] J. Rasley, B. Stephens, C. Dixon, E. Rozner, W. Felter, K. Agarwal, J. Carter, and R. Fonseca. “Planck: Millisecond-scale monitoring and control for commodity networks.” In ACM SIGCOMM, 2014.

[15] A. Sivaraman, S. Subramanian, M. Alizadeh, S. Chole, S.-T. Chuang, A. Agrawal, H. Balakrishnan, T. Edsall, S. Katti, and N. McKeown. “Programmable packet scheduling at line rate.” In ACM SIGCOMM, 2016.

[16] Vibhaalakshmi Sivaraman, Srinivas Narayana, Ori Rottenstreich, S. Muthukrishnan, and Jennifer Rexford. “Heavy-Hitter Detection Entirely in the Data Plane.” In SOSR 2017.

[17] R. Schweller, A. Gupta, E. Parsons, and Y. Chen. “Reversible sketches for efficient and accurate change detection over network data streams.” In ACM IMC, 2004.

[18] Tong Yang, Jie Jiang, Peng Liu, Qun Huang, Junzhi Gong, Yang Zhou, Rui Miao, Xiaoming Li, and Steve Uhlig. “Elastic Sketch: Adaptive and Fast Network-wide Measurements”, In SIGCOMM 2018.

Options for Low-GWP Refrigerants in Small Air-Conditioning Systems[†]

Mark O. McLinden, Andrei F. Kazakov, J. Steven Brown, Riccardo Brignoli, Ian H. Bell, Piotr A. Domanski

Abstract

We summarize a systematic examination of possible low-GWP (global warming potential) replacements for the HFC (hydrofluorocarbon) refrigerants currently used in small air-conditioning systems. The methodology identified the optimal thermodynamic parameters for a refrigerant; this was based on an ideal-cycle analysis, which indicated a tradeoff between efficiency (COP) and refrigeration capacity depending largely on the refrigerant critical temperature. A more realistic analysis, which included an optimization of the heat exchangers, however, revealed that there was a maximum in COP at a relatively high refrigeration capacity, corresponding to refrigerants with a relatively low critical temperature and operating at moderately high pressures. A search in an exhaustive chemical database for fluids having the identified thermodynamic parameters as well as acceptable chemical stability, toxicity, and GWP identified a list of 29 candidate refrigerants. But none of these are a direct, nonflammable, low-GWP replacement for the R-410A currently used in the majority of small air-conditioning systems.

Introduction

The Kigali Amendment to the Montreal Protocol, which phases down the production and consumption of hydrofluorocarbon (HFC) refrigerants, entered into force January 1, 2019. But even before this milestone was reached numerous regional and national regulations concerning the use of HFCs were already in place. For example, the "F-gas" regulations in the European Union[1] mandate maximum values of GWP (global warming potential) for refrigerants in various applications, as well as specifying training and reporting requirements. In the United States, the Significant New Alternatives Program (SNAP) of the U.S. Environmental Protection Agency[2] prohibits or allows specific refrigerants in various applications. The need to identify alternative fluids is obvious.

This paper starts with a brief perspective on the environmental impacts of the HFCs and other refrigerants. The main discussion considers options for low-GWP fluids. First, the methodology of our search for replacement fluids is summarized, and the resulting candidate fluids are presented. Some of the limitations and tradeoffs inherent with replacement fluids are discussed. Finally, brief conclusions are presented. Much of the present paper is drawn directly from our previous works[3,4,5,6,7,8,9,10,11], and the reader is referred to those references for complete details.

Background—Environmental Properties of Refrigerants

Emissions of HFCs currently account for only a small fraction of anthropogenic climate change—about 1 % of the total radiative forcing due to all greenhouse gases, including CO_2, CH_4, and N_2O, according to the 2018 Scientific Assessment of Ozone Depletion.[12] The Kigali Amendment and other regulations are driven by estimated large increases in future emissions predicted under "business as usual" scenarios as laid out by Velders *et al.*[13,14] With the HFC phasedown of the Kigali Amendment, Velders[15] estimates that the surface temperature increase from HFCs will be limited to 0.06 K by the end of the century compared to (0.3 to 0.5) K with "business as usual."

The environmental characteristics of selected refrigerants, as compiled by the WMO,[12] are listed in Table 1. The GWP parameter was first presented by the Intergovernmental Panel on Climate Change[16] in 1990 as a simple

[†] Presented at the Annual General Meeting of the Institute of Refrigeration, London, 6 November 2019

McLinden, Mark O.; Kazakov, Andrei F.; Brown, J; Brignoli, Riccardo; Bell, Ian; Domanski, Piotr A. "Options for Low-GWP Refrigerants."
Paper presented at Institute of Refrigeration, London, UK. November 06, 2019 - November 06, 2019.

metric for comparing the impact of different gases on the climate system. The GWP of a gas is relative to that of carbon dioxide; furthermore, it is calculated over some time interval, or "time horizon." A time horizon of 100 years is most commonly used (including in regulations) and is referred to as GWP_{100}. Other time horizons are sometimes used, and Table 1 also lists the GWP_{20}—the value for a 20-year time horizon. For long-lived gases, such as R-12, the GWP_{100} and GWP_{20} are similar. For gases with lifetimes on the order of a few decades or less (*e.g.*, the HFCs), however, the GWP_{20} is substantially higher than the GWP_{100} because the climate impact of such gases is concentrated in the initial years following their release into the atmosphere. When calculating over a longer time horizon, the impact in the later years is small, and the GWP_{100} basically "averages out" the impacts. This indicates that the short-term climate impact of the HFCs can be understated by their GWP_{100}. But this also underlies an opportunity for a positive impact on the climate by mid-century by phasing down the HFCs: CO_2 is long-lived in the atmosphere, and so the benefits of any reduction in its emissions are damped by its existing concentration in the atmosphere. A phase-down of HFC emissions, on the other hand, has much more immediate benefits.

Table 1. Environmental characteristics of selected refrigerants.[†]

R-number	Formula	T_{NBP} (˚C)	Lifetime (years)	ODP	GWP_{20}	GWP_{100}
CFCs and HCFCs						
R-12	CCl_2F_2	–29.8	102	0.73–0.81	10800	10300
R-22	$CHClF_2$	–40.8	11.9	0.024–0.034	5310	1780
R-123	$CHCl_2CF_3$	27.8	1.3	0.01	290	80
HFCs						
R-32	CH_2F_2	–51.7	5.4	0	2530	705
R-125	CHF_2CF_3	–48.1	30	0	6280	3450
R-134a	CH_2FCF_3	–26.1	14	0	3810	1360
R-152a	CH_3CHF_2	–24.0	1.6	0	545	148
HFOs and HCFOs						
R-1233zd(E)	$CF_3CH{=}CHCl$	18.3	0.071	<0.0004	13.5	3.7
R-1234yf	$CF_3CF{=}CH_2$	–29.5	0.029	0	1	<1
R-1234ze(E)	$CF_3CH{=}CFH$	–19.0	0.045	0	4	<1
Natural fluids						
R-290	$CH_3CH_2CH_3$ (propane)	–42.1	0.041	0	<1	<1
R-717	NH_3 (ammonia)	–33.3	"few days"	0	<1	<1
R-744	CO_2 (carbon dioxide)	–56.6[‡]	–	0	1.00	1.00

[†] Environmental data from WMO[12]; T_{NBP} from REFPROP[17]; table from ref[10].
[‡] Triple-point temperature

While the many of the HFCs do have high values of GWP it must be kept in mind that they were developed and implemented to replace substances with high ozone-depletion potential (ODP), the so-called ozone-depleting substances or ODSs, primarily the CFCs and HCFCs. Many of the CFCs and HCFCs also have high values of GWP (see Table 1); for example, the GWP_{100} of the most common refrigerant in automotive systems was reduced by a factor of 7.5 (10300 to 1360) when R-12 was replaced with R-134a. Thus, the Montreal Protocol, which was originally targeted at stratospheric ozone protection, has already had a tremendously positive effect on climate, as reported by Velders *et al.*[18] In that analysis, Velders *et al.* consider a scenario in which the 1987 Montreal Protocol (and subsequent amendments) was not implemented. Without any action, the radiative forcing (*i.e.*, impact on climate) of the ODSs would have been comparable to CO_2 emissions by 2010. Instead, the actual yearly GWP-weighted emissions of the ODSs peaked in 1988 at an equivalent equal to 43 % of that of CO_2. By

2010, the Montreal Protocol had avoided emissions of approximately 10 Gt CO_2[eq] (*i.e.*, emissions with a warming impact equivalent to 10 gigatonnes of CO_2) per year; by comparison, global CO_2 emissions from fossil fuels and industrial processes were 32 Gt per year in 2010.[19] Put another way, the Montreal Protocol has delayed climate change by 7–12 years, according to Velders *et al.*[18]

Thus, by rapidly phasing out the CFCs and other ODSs in the 1990s, the refrigeration industry has already avoided significant climate impacts, and with the HFC phasedown called out in the Kigali Amendment to the Montreal Protocol the industry will make further substantial contributions.

The Search for Replacement Fluids

The present paper summarizes the results of a comprehensive search for the best single-component, low-GWP replacement fluids for use in small air-conditioning systems, as presented by McLinden *et al*.[9] We searched for suitable replacement fluids by applying thermodynamic and environmental screening criteria to a comprehensive chemical database. The fluids passing these screens were then simulated in an air-conditioning system, with the calculated volumetric refrigeration capacity and COP (*i.e.*, energy efficiency) serving as additional screens.

A refrigerant is the essential working fluid in a vapor-compression refrigeration cycle; it absorbs heat at a relatively low temperature in the evaporator (*e.g.*, the cooling coil in an air conditioner) and releases it at a higher temperature in the condenser (*e.g.*, the outside coil). To identify replacement refrigerants one must first consider the characteristics required of any refrigerant and then also consider the optimal properties for the particular application of interest.

General requirements of a refrigerant. The requirements of a replacement refrigerant have been considered by numerous authors including McLinden and Didion,[20] Calm and Didion,[21] and Kujak and Schultz.[22] Always at the top of the list is the need for chemical stability within the sealed refrigeration system; next are health and safety considerations. Existing safety codes (*e.g.*, ASHRAE Standard 15[23]) require nonflammable refrigerants for many applications, but that requirement is being reconsidered. Environmental considerations (*e.g.*, ODP and GWP) are obviously important—they are the reason new fluids are currently under consideration. Practical requirements such as cost and materials compatibility also factor in.

A refrigerant must also possess certain thermodynamic characteristics for it to function. Since a refrigerant absorbs and releases heat primarily through evaporation and condensation, a high latent heat of vaporization might seem to be desirable. A boiling-point temperature low enough that air will not leak into a system is desirable; but the boiling point should not be so low that the system pressures are "too high." A high discharge temperature upon compression will lower the efficiency of the cycle, but conversely, "wet compression" must be avoided in most types of compressors.

The "Exploration of Thermodynamic Space." But how are these seemingly disparate thermodynamic characteristics related, and how is one to identify a refrigerant satisfying them? McLinden *et al.*[4] and Domanski *et al.*[5] approached this problem by defining fluids in terms of a small number of fundamental thermodynamic characteristics and then searching for optimal values of those parameters. Thus, they considered the full range of possible thermodynamic behaviour, rather than scanning a finite number of known fluids (which would reveal no new fluids). They termed this approach the "exploration of thermodynamic space."

The fluid thermodynamic properties were modeled by the "extended corresponding states" (ECS) approach laid out by Huber and Ely.[24] Corresponding states is the observation that the properties of fluids are similar when scaled by their critical point parameters, T_{crit} and p_{crit}, where the critical point is the state where the saturated liquid and saturated vapor approach one another. Key to this method is a "reference fluid" to which the properties of the unknown fluid are scaled. This method is "extended" by additional parameters; the most significant were the heat capacity of the vapor ($C_p{}^0$) and the acentric factor, ω, which is related to the slope of the vapor pressure curve. The range of the parameters considered is given in Table 2. The critical temperature is not as familiar as the normal boiling point temperature, T_{NBP}, but the two are related, and here both are used here. (Other parameters were considered, but they were found to be of minor importance.[5])

Table 2. Fluid parameters varied in the optimization runs and their ranges.

Parameter	Range
reference fluid	propane –or– R-32
T_{crit}/K	305 – 650
p_{crit}/MPa	2.0 – 12.0
ω	0.0 – 0.6
$C_p{}^0$(300 K)/J·mol^{-1}·K^{-1}	20.8 – 300

Optimum Thermodynamic Characteristics. Any optimization requires objective function(s), and here the coefficient of performance (COP) and volumetric capacity (Q_{vol}) were selected for a cycle operating between an evaporation temperature of 10 ˚C and a condensation temperature of 40 ˚C. The hypothetical fluids were simulated in three cycles (shown in Figure 1): (a) the simple (basic) four-component vapor-compression cycle, (b) a cycle with a liquid-line/suction-line heat exchanger (LL/SL HX), and (c) a two-stage economizer cycle. All three cycles were modelled assuming isentropic compression, no pressure drop in the heat exchangers, and saturated liquid and vapor exiting the condenser and evaporator, respectively. Sets of thermodynamic parameters within the ranges defined in Table 2 defined a series of hypothetical fluids. By varying these parameters according to an evolutionary algorithm, optimal values were determined; see McLinden *et al*.[4] and Domanski *et al*.[5] for details.

This exploration of thermodynamic space indicated a trade-off between efficiency and volumetric capacity. Refrigerants with a high critical temperature gave high efficiency, but low capacity; fluids with a relatively low T_{crit} resulted in the converse. A critical pressure at the upper limit of the range resulted in both higher efficiency and increased capacity, while an acentric factor near the lower limit was optimal. The optimum value of the vapor heat capacity varied with the cycle; a relatively low value was best for the simple vapor-compression cycle, while a higher value was optimal for a cycle with internal heat exchange between the condenser outlet and compressor inlet. These results are presented in Figure 2; a clear efficiency-versus-capacity tradeoff is seen for all three cycles. Figure 2 also shows the "Pareto front" comprising the set of hypothetical fluids having thermodynamic parameters offering the highest COP for a given Q_{vol} and vice-versa; this defines the upper limit of performance that is allowed by thermodynamics. (The inverse of COP and Q_{vol} are plotted to yield a minimization, which is the convention for this type of problem.) For selected "real" refrigerants, the cycle with the LL/SL heat exchanger (Figure 2(b)) shows a much smoother variation of COP versus Q_{vol} compared to the simple cycle (Figure 2(a)); some refrigerants, notably ammonia and R-32, have a lower COP with the LL/SL HX. The economizer cycle (Figure 2(c)) shows a higher COP for both the Pareto front and the current refrigerants compared to the simple vapor-compression cycle. The relative benefit of the LL/SL heat exchanger cycle versus the economizer cycle compared to the simple cycle varied with the fluid.

Database Screenings. Having defined a desired set of thermodynamic parameters, we next set out to identify fluids having those characteristics. Our search relied on screening a comprehensive database of molecules by applying filters representing different refrigerant selection criteria. The search was carried out in the PubChem database—a listing with more than 60 million chemical structures.[25] A first screening of this database is described by Kazakov *et al*.[3]; we summarize here a second screening.[9] All current refrigerants are small molecules, and McLinden[26] provides a thermodynamic basis for this. Thus, we limited our search to molecules with 18 or fewer atoms and comprising only the elements C, H, F, Cl, Br, O, N, or S. The choice of elements follows the observation by Midgley[27] that only a small portion of the periodic table would form compounds

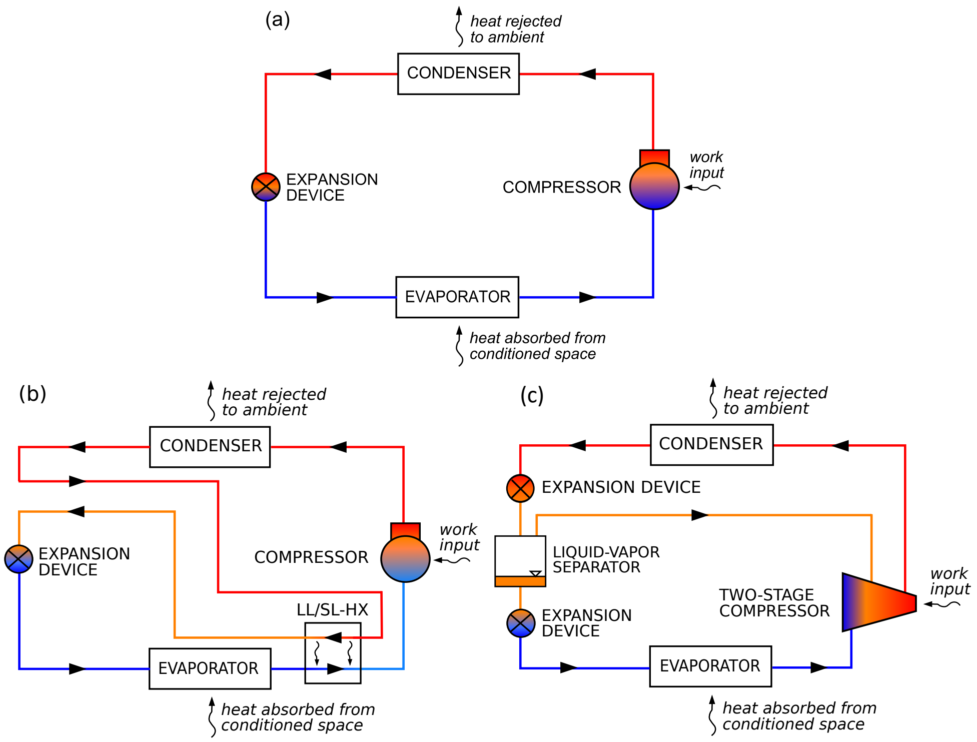


Figure 1. Cycles simulated: (a) simple vapor-compression cycle; (b) cycle with 100 % effective LL/SL heat exchanger; (c) economizer cycle; figure from ref [9].

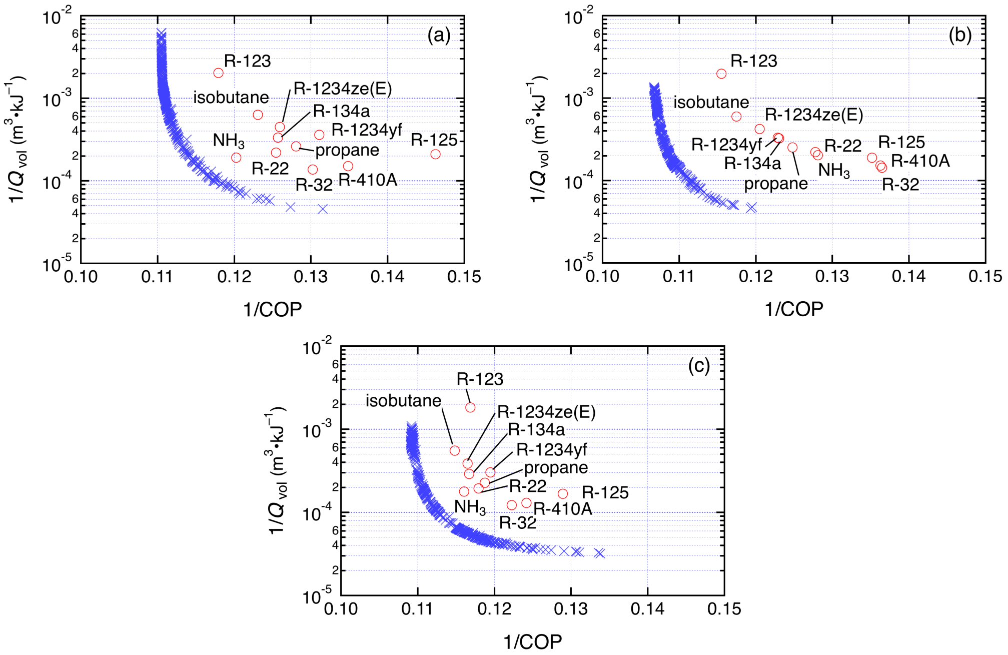


Figure 2. Pareto front (×) and selected current refrigerants (○) for different cycle options: (a) simple vapor-compression cycle; (b) cycle with 100 % effective LL/SL heat exchanger; (c) economizer cycle; figure from ref.[4].

volatile enough to serve as refrigerants. Despite their ability to deplete stratospheric ozone, chlorine and bromine were included since molecules including Cl or Br might have a negligible ODP and might be acceptable if they had a very short atmospheric lifetime. These restrictions on elements and molecular size resulted in 184 000 molecules to be considered further.

Further screens for 320 K < T_{crit} < 420 K and GWP_{100} < 1000 yielded 138 fluids. The PubChem database does not provide T_{crit} and GWP_{100} values for the vast majority of the compounds, so they were estimated using methods based solely on molecular structure; these estimations constituted a major effort of the project.[28,3,4,29] The limits on critical temperature correspond to fluids usable in small AC systems, with an allowance for the uncertainty in the estimated values of T_{crit}. While refrigerants with values of GWP as low as possible are obviously desirable, fluids with GWP_{100} < 750 are, for example, permitted under E.U. regulations in AC systems with less than 3 kg of refrigerant.[30] (The full list of 138 fluids is given in the Supplementary Information of McLinden *et al*.[9], which also lists the estimated T_{crit} and GWP_{100} for each fluid.)

The next screens were for chemical stability and toxicity. Compounds with generally unstable functional groups were dropped from further consideration. For example, peroxides (compounds with the –O–O– group) are unstable. Ketenes (compounds with the –C=C=O group) are generally very reactive, and three such compounds were dropped. Allenes have the –C=C=C– group and are characterized as "difficult to prepare and very reactive."[31] Compounds with a carbon-carbon triple bond are generally less stable than those with a double bond; for example, fluoroethyne (FC≡CH) is described as "treacherously explosive in the liquid state."[32] There are exceptions, however, and trifluoropropyne was retained.

Attempts to automate the screening of toxicity were not successful. Fortunately, at this point, the number of compounds was sufficiently small to allow a "manual" examination of toxicity data. We considered published toxicity data, where available, making use of a variety of sources, including safety standards, compilations of toxic industrial chemicals, regulatory filings, and safety data sheets of chemical manufacturers. We also dropped compounds with two specific groups. Molecules that included the $=CF_2$ group were deemed "not viable candidates" on the basis of Lindley and Noakes[33] who discuss the "$=CF_2$ structural alert" in regards to R-1225zc ($CF_3CH=CF_2$); this is the observation that the $=CF_2$ group has a high reactivity which is often associated with toxic effects. The presence of a $=CF_2$ group does not assure that a molecule is toxic, but we are aware of only one possible counterexample of R-1123 ($CHF=CF_2$), which has an acute toxicity similar to that of the commercialized refrigerant R-1234ze(E).[34] (The chronic toxicity of R-1123 has not been reported in the public literature.) The absence of a $=CF_2$ group does not, however, imply that a molecule is of low toxicity. Fluids having the –OF group were also dropped. The –OF group is analogous to the –OH group that defines an alcohol. The bond dissociation energy of the O–F bond, however, is less than one-half that of the O–H bond in an alcohol,[35] and the fluorine would likely be reactive with water, forming HF (hydrofluoric acid—a highly toxic compound).

Performance in the Ideal Vapor-Compression Cycle. The 138 candidates identified in the database screening were simulated in the simple (ideal) vapor-compression cycle. For the representation of refrigerant properties we used detailed equations of state (EOS) implemented in the NIST REFPROP database[36] where available. However, for a majority of fluids we used the extended corresponding states (ECS) model,[4,24] as discussed above. This screening proceeded in two rounds. The first round of cycle simulations made use of the theoretical CYCLE_D model[37] and provided a first estimate of volumetric capacity and COP.[7] These simulations assumed an ideal cycle with 100 % compressor efficiency and no pressure drops. At this stage, we dropped fluids with a volumetric capacity less than one-third that of R-410A or a COP < 5. (For R-410A in the ideal cycle Q_{vol} = 6.62 $\mathrm{MJ{\cdot}m^{-3}}$ and COP = 7.41.)

Candidate Fluids

The database search combined with the performance simulation in the ideal cycle resulted in a list of 27 fluids with the assertion that it would be "highly unlikely that any better-performing fluids will be found."[9] The list already included several fluids that failed the stated screening criteria, but were added because of commercial interest. These included R-1123, which was identified as unstable by Tanaka *et al.*[38], but a blend with R-32 may be stable.[39] Have any additional fluids been identified in the intervening time? Domanski *et al.*[8] added R-1132a to the list. This fluid was among the 138 fluids passing the original screening but was rejected based on an indication of high toxicity[40]; later data resulted in a "class A" toxicity rating.[41]

The iodine-containing CF_3I has generated considerable interest recently as a low-GWP option that might be used to suppress the flammability of blends (see, for example, McGowan[42]). CF_3I was recently classified as "A1" (low-toxicity, nonflammable) under ASHRAE Standard 34 and designated as R-13I1.[43] This fluid was considered early on by Kazakov *et al.*[3] The C–I bond is very weak and is readily photolyzed by UV light; thus, it did not fit within the estimation scheme for GWP developed by Kazakov. CF_3I and other iodine-containing compounds were investigated in the 1990s during the search for CFC replacements (see Calm[44] and Nimitz and Lankford[45]) and again in the early 2000s in the program sponsored by the Society of Automotive Engineers to identify replacements for R-134a in automotive air-conditioning systems.[46] Interest waned in the iodine-containing compounds on both occasions because of stability and toxicity concerns. Thus, iodine was not included in the list of eight elements making up the fluids in the database searches of Kazakov *et al* [3] and McLinden *et al.*, [6,9] but we include it here.

The final list of 29 candidates is given in Table 3. This list is a subset of the 138 candidates having 320 K < T_{crit} < 420 K and GWP_{100} < 1000, with the deletion of those that have low Q_{vol}, low COP, or are unstable or toxic. The list comprises four hydrocarbons and the closely related dimethylether; five fluorinated alkanes (*i.e.*, HFCs); ten fluorinated alkenes and an alkyne; two fluorinated oxygen-containing molecules; three fluorinated nitrogen or sulfur compounds; CF_3I; and two inorganic molecules (ammonia or NH_3 and carbon dioxide). The list includes a small number of novel molecules that have not been previously considered as refrigerants (at least publicly), but a majority of the fluids are well known, including ammonia (R-717) and propane (R-290), or are the focus of current research in the refrigeration industry, *i.e.*, the fluorinated alkenes (also known as hydrofluoroolefins or HFOs).

Refrigerant blends are currently in common use, and the fluids in Table 3 also constitute the components of future blends. Several of the fluids have low critical temperatures and would operate in a transcritical cycle for our investigated conditions; thus, they were not simulated. These fluids might, however, be useful as a component in a refrigerant blend, and they were included in Table 3 for this reason. The table also includes an additional four current HFCs and HCFCs for comparison purposes. Despite their high values of GWP_{100} R-134a and R-125 might also find use as blend components. Some of the reasons behind this paucity of candidates are discussed below.

Simulations in an Optimized Cycle. The second round of simulations made use of a more advanced "optimized" cycle model that provided a more realistic representation of an air conditioner employing typical forced-convection, air-to-refrigerant heat exchangers, which were optimized for a particular refrigerant.[47] In this type of heat exchanger, the refrigerant undergoes a phase change as it flows down the inside of a tube and exchanges heat with air on the outside of the tube. Specifically, the model accounted for the effect of optimized refrigerant mass flux, which enhances the refrigerant heat transfer coefficient at an acceptable penalty of the pressure drop. These simulations included the effects of the transport properties (thermal conductivity and viscosity) on the heat transfer coefficient and pressure drop, which were not included in the ideal cycle simulations. The simulation model maintained the same heat flux in the evaporator through all simulations, which is a prerequisite for a fair rating of competing refrigerants.[48] The isentropic efficiency of the compressor was a function of the refrigerant properties and averaged 70 %.[49] Here, the relative ranking of fluids differs from a ranking based only on thermodynamic properties; it is, however, more representative of a fluid's performance in an AC system in commercial production, which would be optimized for the refrigerant being used.

The COP and Q_{vol} of the candidate fluids, based on the optimized model, are presented in Table 3 and Figure 3. Unlike the COP versus Q_{vol} tradeoff observed for the ideal analysis (Figure 2), the results of the optimized cycle simulations in the simple cycle (Figure 3) show a maximum in COP corresponding to Q_{vol} of approximately 60 %

Table 3. COP and volumetric capacity of selected low-GWP fluids and current HFC and HCFC fluids. Results are presented for the basic, liquid-line/suction-line heat exchanger (LL/SL), and economizer (Econ.) cycles. Values are for the "optimized" cycle model and are relative to the performance of R-410A in the basic cycle. GWP_{100} are from WMO[12] or E.U. regulation[30] unless noted. The fluids are grouped by chemical class and, within classes, listed in order of increasing critical temperature; table adapted from ref [9].

IUPAC Name	Structure	ASHRAE Designation	GWP_{100}	$COP/COP_{R\text{-}410A}$*			$Q_{vol}/Q_{vol,R\text{-}410A}$*		
				Basic	LL/SL	Econ.	Basic	LL/SL	Econ.
Hydrocarbons and dimethylether									
ethane	CH_3-CH_3	R-170	6	§					
propene (propylene)	CH_2=CH-CH_3	R-1270	<1	1.033	1.053	1.073	0.689	0.694	0.770
propane	CH_3-CH_2-CH_3	R-290	<1	1.014	1.042	1.058	0.571	0.579	0.640
methoxymethane (dimethylether)	CH_3-O-CH_3	R-E170	1	0.996	1.002	1.035	0.392	0.389	0.427
cyclopropane	-CH_2-CH_2-CH_2-	R-C270	86†	1.018	1.021	1.045	0.472	0.467	0.510
Fluorinated alkanes (HFCs)									
fluoromethane	CH_3F	R-41	116	§					
difluoromethane	CH_2F_2	R-32	705	1.038	1.026	1.070	1.084	1.057	1.191
fluoroethane	CH_2F-CH_3	R-161	6	1.026	1.031	1.062	0.601	0.594	0.658
1,1-difluoroethane	CHF_2-CH_3	R-152a	148	0.981	0.989	1.022	0.399	0.396	0.435
1,1,2,2-tetrafluoroethane	CHF_2-CHF_2	R-134	1135	0.967	0.991	1.024	0.348	0.352	0.385
Fluorinated alkenes (HFOs) and alkyne									
1,1-difluoroethene	CF_2=CH_2	R-1132a	<1	§					
fluoroethene	CHF=CH_2	R-1141	<1	0.968	0.977	1.014	1.346	1.336	1.547
1,1,2-trifluoroethene	CF_2=CHF	R-1123	3†	0.956	0.988	1.014	1.054	1.074	1.230
3,3,3-trifluoroprop-1-yne	CF_3-C≡CH	n.a.	1.4†	0.988	1.023	1.042	0.545	0.557	0.616
2,3,3,3-tetrafluoroprop-1-ene	CH_2=CF-CF_3	R-1234yf	<1	0.954	1.006	1.020	0.414	0.431	0.474
(E)-1,2-difluoroethene	CHF=CHF	R-1132(E)	1†	1.016	1.019	1.051	0.591	0.585	0.646
3,3,3-trifluoroprop-1-ene	CH_2=CH-CF_3	R-1243zf	<1	0.964	0.997	1.019	0.372	0.379	0.417
1,2-difluoroprop-1-ene‡	CHF=CF-CH_3	R-1252ye‡	2†	0.973	0.996	1.021	0.355	0.358	0.392
(E)-1,3,3,3-tetrafluoroprop-1-ene	CHF=CH-CF_3	R-1234ze(E)	<1	0.939	0.977	1.004	0.320	0.329	0.360
(Z)-1,2,3,3,3-pentafluoro-1-propene	CHF=CF-CF_3	R-1225ye(Z)	<1	0.922	0.972	0.986	0.273	0.285	0.310
1-fluoroprop-1-ene‡	CHF=CH-CH_3	R-1261ze‡	1†	0.975	0.983	1.018	0.353	0.351	0.385
Fluorinated oxygenates									
trifluoro(methoxy)methane	CF_3-O-CH_3	R-E143a	540	0.957	0.992	1.017	0.366	0.374	0.411
2,2,4,5-tetrafluoro-1,3-dioxole	-O-CF_2-O-CF=CF-	n.a.	1†	0.936	0.984	0.998	0.337	0.349	0.376
Fluorinated nitrogen and sulfur compounds									
N,N,1,1-tetrafluormethaneamine	CHF_2-NF_2	n.a.	20†	0.965	1.007	1.027	0.807	0.831	0.937
difluoromethanethiol	CHF_2-SH	n.a.	1†	1.010	1.019	1.054	0.582	0.580	0.642
trifluoromethanethiol	CF_3-SH	n.a.	1†	0.977	0.997	1.026	0.418	0.421	0.464
Iodine compound									
trifluoroiodomethane	CF_3I	R-13I1	<1	0.913	0.927		0.310	0.310	
Inorganic compounds									
carbon dioxide	CO_2	R-744	1.00	§					
ammonia	NH_3	R-717	<1	1.055	1.028	1.080	0.746	0.721	0.791
Current HFCs and HCFCs									
pentafluoroethane	CF_3-CHF_2	R-125	3450	0.913	0.979	0.995	0.746	0.784	0.889
R-32/125 (50.0/50.0)	blend	R-410A	2078	1.000	1.012	1.049	1.000	0.997	1.130
chlorodifluoromethane	$CHClF_2$	R-22	1780	1.007	1.008	1.043	0.666	0.658	0.732
1,1,1,2-tetrafluoroethane	CF_3-CH_2F	R-134a	1360	0.968	0.993	1.027	0.433	0.439	0.485

*Values are relative to those for R-410A in the basic cycle; $COP_{R\text{-}410A} = 5.35$ and $Q_{vol,R\text{-}410A} = 6.93\ MJ{\cdot}m^{-3}$.
†Value estimated by the method of Kazakov *et al.*[3]
‡This fluid has cis (Z) and trans (E) isomers; the predicted values of both were the same.
§Fluid would be near-critical or supercritical in the condenser and was not simulated.

to 110 % that of R-410A. Although there is considerable scatter, a polynomial curve fitted to the fluids shown in Figure 3 indicates the general trend. Relative to fluids with low values of Q_{vol}, the high-Q_{vol} fluids have lower values of T_{crit} and operate at higher pressures; the result is that the cycle operates near the critical point and suffers increased irreversibilities in the expansion process. This effect applies to both the ideal and more-detailed analyses. However, the ideal analysis neglects the fact that the pressure drop in the heat exchangers (condenser and evaporator) extracts a smaller COP penalty on the high-Q_{vol} (*i.e.*, high-pressure) fluids when the heat exchangers are optimized. An additional effect is that the low-Q_{vol} fluids tend to be more complex molecules. For example, R-32 (one of the best fluids in Figure 3) is based on a single carbon atom, and R-410A is a blend of the single-carbon R-32 and two-carbon R-125. In contrast most of the fluids with $Q_{vol} < 0.4 \cdot Q_{vol,R\text{-}410A}$ are three-carbon compounds; greater molecular complexity is associated with higher values of viscosity, which would increase the pressure drop and lower the COP.

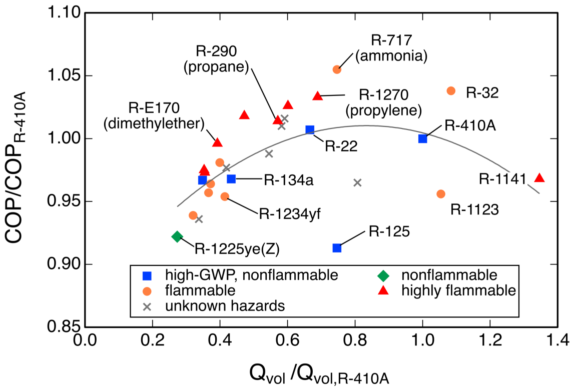


Figure 3. COP and Q_{vol} of selected low-GWP fluids relative to R-410A in the basic vapor-compression cycle including pressure drop and heat-transfer limitations; the curve indicates the general trend; figure adapted from ref [9].

The COP ranged from –7.8 % to +5.5 % relative to that of R-410A in the basic vapor-compression cycle. Ammonia showed the highest COP, better than that for R-410A by 5.5 %. Beyond ammonia, which is toxic, mildly flammable, and presents materials compatibility issues, the COPs of R-32, propene (R-1270), R-161, R-1132(E), propane (R-290), cyclopropane (R-C270) and difluoromethanethiol are also above the R-410A baseline. The COPs of the remaining fluids are lower. Mildly flammable R-32 has a COP and Q_{vol} higher than that of R-410A, but this advantage comes with a GWP_{100} of 705. R-134 and R-E143a have GWP_{100} values of 1135 and 540, respectively. Three fluids have GWPs within the 80 to 150 range, and GWPs for the remaining fluids do not exceed 20. Except for R-32, R-1123, and R-1141 the listed fluids have Q_{vol} lower than R-410A and would thus require a larger compressor—by at least 25 % and, for a majority of the candidates, more than twice as large—to provide the same capacity as R-410A. Table 3 does not provide COP and Q_{vol} for carbon dioxide, ethane, R-41, and R-1132a because their T_{crit} are low and they may require a different (*i.e.*, transcritical) cycle, depending on operating conditions. We list them because they may be suitable as a component of a blend. In general, fluids with a low T_{crit} (corresponding to high Q_{vol}) suffer performance degradation at high ambient temperatures.

The results for the LL/SL-HX and economizer cycles are qualitatively similar to the basic cycle and are listed in Table 3. The LL/SL-HX cycle provides a performance benefit to fluids with a high vapor heat capacity and degrades the performance of fluids with a low vapor heat capacity (which are best performers in the basic cycle). Consequently, the spread of COP values is smaller than that shown in Figure 3. The economizer cycle increases the COP for all refrigerants, although the increase is larger for the fluids having a high vapor heat capacity.

Refrigerant Blends. Given the lack of an obvious low-GWP, nonflammable, pure-fluid candidate with a volumetric capacity similar to R-410A there is considerable interest in refrigerant blends; see, for example, Schultz[50] and Spletzer *et al.*[51] Mixing high-GWP, but non-flammable fluid(s) with low-GWP, but flammable fluid(s) can yield a mildly flammable refrigerant with a GWP_{100} on the order of 500 or, with a different composition, a non-flammable fluid with a higher GWP_{100}.

Bell, *et al.*[11] considered the problem of finding a nonflammable replacement for R-134a in an air-conditioning application. They considered blends composed from a slate of 13 candidate fluids with a range of pressure, flammability, and GWP values that might produce a blend with the desired characteristics of a R-134a replacement. The blend components included hydrofluoroolefins (HFOs), which have very low GWP values (≈ 1 relative to CO_2), but that are mildly flammable; hydrofluorocarbons (HFCs) with moderate-to-high GWP values that were nonflammable and thus, might serve to suppress the flammability of a blend; additional mildly flammable HFCs; and carbon dioxide (CO_2), which is nonflammable with GWP ≡ 1, but which would raise the working pressure of a blend and may produce a large two-phase temperature glide in the heat exchangers. All the selected fluids were of low toxicity, *i.e.*, having an "A" classification under ASHRAE Standard 34.[52] Additional considerations were the commercial availability of the fluid and the availability of property data (in the form of an accurate equation of state), so that cycle simulations could be carried out with some measure of confidence. The blend components were the HFCs: R-134a, R-227ea, R-125, R-143a, R-32, R-152a, R-134, and R-41; the HFOs: R-1234yf, R-1234ze(E), R-1234ze(Z), and R-1243zf; and R-744 (CO_2). Hydrocarbons were not included because the objective was to find a nonflammable blend, and a blend containing only a few percent of a hydrocarbon would be flammable.

A simplified cycle model was used, but it did include the effects of compressor efficiency, pressure drop in the condenser and evaporator, subcooling at the condenser outlet, and superheat at the evaporator outlet. The basic cycle conditions were an evaporator dew-point temperature of 10 ˚C and condenser bubble-point temperature of 40˚C. In contrast to the exploration of thermodynamic space, which employed an optimization based on an evolutionary algorithm, Bell *et al.*[11] simulated all possible combinations of the 13 components, for a total of 100 387 binary and ternary mixtures to be evaluated. All possible four-component mixtures were also considered, for an additional 1.4 million evaluations (although none of the four-component blends were superior to the binary or ternary blends). The flammability of the blends was estimated with a new scheme developed by Linteris *et al*.[53] which yields a "normalized flammability index" $\overline{\Pi}$, which varies from $\overline{\Pi} = 100$ for highly-flammable hydrocarbons (*i.e.*, containing no fluorine); $\overline{\Pi} = 0$ at the limit of flammability (*i.e.*, the boundary of the ASHRAE "1" (nonflammable) and "2L" (mildly flammable) classes); and $\overline{\Pi} < 0$ for nonflammable fluids. Note that $\overline{\Pi}$ is based on the ASTM E-681 test method[54] specified in the ASHRAE standard,[52] and its estimates of flammability versus nonflammability may differ for other test methods.

Figure 4 provides an overview of the results for the binary and ternary blends. Here the COP is plotted versus GWP_{100} and sorted into four "bins" of estimated flammability; each dot represents one blend composition. Starting at the right-most bin, which includes blends which are predicted to be clearly nonflammable, all of the blends have substantially lower COP compared to the R-134a baseline, indicated with the red dashed line. (Note that only blends with GWP_{100} less than or equal to R-134a are plotted.) The next bin to the left shows the blends that are predicted to be nonflammble, but near the flammability limit (*i.e.*, $-10 < \overline{\Pi} < 0$). Here, only blends with $GWP_{100} > 535$ have COPs within a few percent of R-134a, and the upper bound of the COPs increases with GWP_{100}, approaching the COP of R-134a at high values of GWP_{100}. The second bin from the left contains blends that are estimated to be slightly flammable (*i.e.*, $0 < \overline{\Pi} < 10$). (For comparison, $\overline{\Pi} = 4.8$ for R-1234yf and $\overline{\Pi} = -10.5$ for R-134a.) This bin "fills in" the upper-left quadrant of the bin to the right, that is, blends with $GWP_{100} < 535$ and COP approaching that of R-134a. Only the left-most bin, comprising blends that are somewhat more flammable, but still estimated to be in the "2L" flammability classification, contains fluids which have both low values of GWP_{100} and COPs equal to, or greater than, the R-134a baseline.

The yellow boxes in the middle two bins indicate the relatively small number of blends that would be of interest in the search for a nonflammable R-134a replacement, *i.e.*, those with COPs similar to that of R-134a and that are nonflammable or only mildly flammable. The blends identified in this study are applicable only for its specific objectives, but it hints at the difficulty of finding a suitable blend.

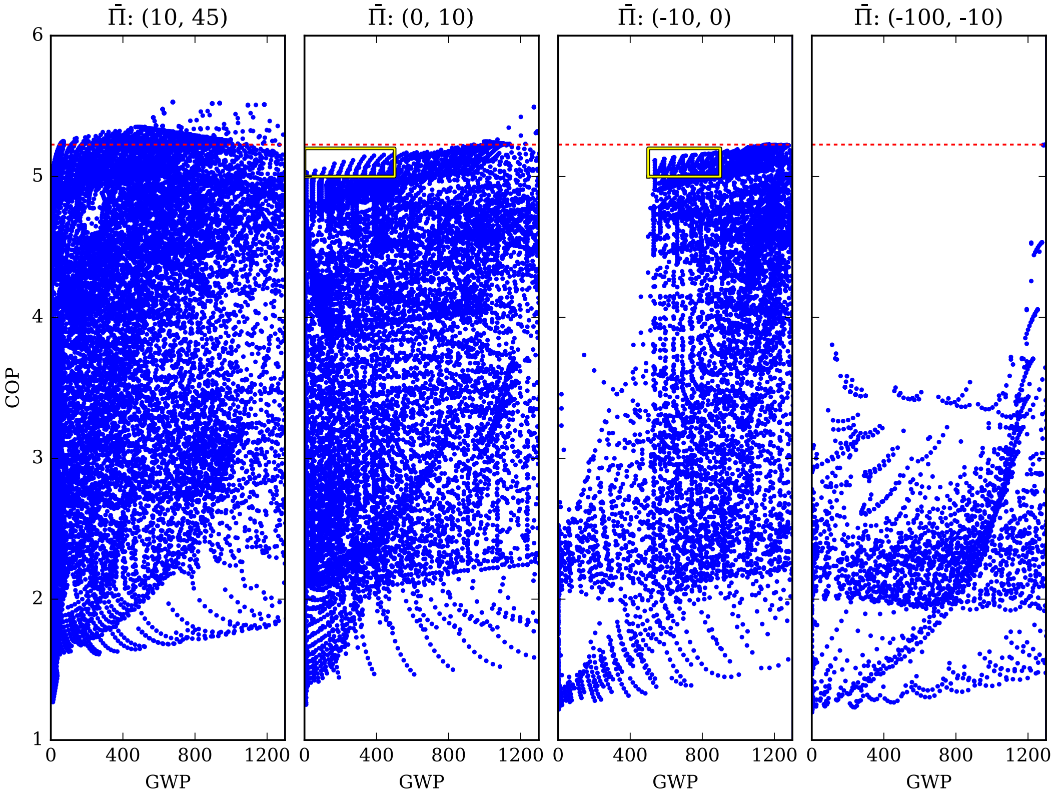


Figure 4. Overview of cycle simulation results sorted into bins of estimated flammability. The flammability decreases moving from left to right; the two right-most bins are predicted to be non-flammable. The red dashed line indicates the COP of the R-134a baseline system, and the yellow boxes indicate the blends of greatest interest. Figure from Bell *et al*.[11]

Limitations and Tradeoffs

When Midgley famously introduced R-12 at the 1930 meeting of the American Chemical Society by inhaling a lungful of the refrigerant and using it to extinguish a burning candle,[55,56] it seemed that the perfect refrigerant had been found. It was nonflammable, of low toxicity, and was simple and cheap to manufacture. It had good thermodynamic properties, and, in fact, its properties were nearly a perfect match to the simple vapor-compression cycle.[26] But, of course, it was later found to deplete stratospheric ozone and have a high GWP. Even today, R-12 remains the fourth-most powerful greenhouse gas, behind CO_2, methane, and N_2O because of past emissions.[12] Since the mid-1970s the task of identifying suitable refrigerants has faced an increasing list of constraints. And, while these constraints have restricted the choice of fluids (*e.g.*, CFCs and HCFCs are no longer available), they have also opened up interest in other fluids. The phaseout of the ozone-depleting fluids, for example, sparked a resurgence of interest in the long-known natural fluids (*e.g.*, ammonia, hydrocarbons). Likewise, the lack of an obvious low-GWP, nonflammable, high-pressure replacement for some of the HFCs has opened a discussion on how fluids with at least some degree of flammability might be safely implemented in a wider variety of systems.

Constraints on molecular structure. As environmental regulations became ever more restrictive, the constraints on the molecules suitable for use as refrigerants correspondingly increased. When the CFCs were first introduced in the 1930s, the constraints were to find a molecule that was nonflammable, of low toxicity, cheap to manufacture and with suitable thermodynamic characteristics. The halogenated alkanes can be represented on

McLinden, Mark O.; Kazakov, Andrei F.; Brown, J; Brignoli, Riccardo; Bell, Ian; Domanski, Piotr A. "Options for Low-GWP Refrigerants." Paper presented at Institute of Refrigeration, London, UK. November 06, 2019 - November 06, 2019.

a triangular grid, with a separate grid for molecules based on one, two, etc. carbons. The base hydrocarbon (methane, ethane, etc.) is at the top of the triangle, and the bottom corners represent molecules with the hydrogens fully substituted with either chlorine or fluorine. With no environmental constraints, CFCs and HCFCs based on methane (one carbon), namely R-12 and R-22, were available to meet a variety of applications, as shown in Figure 5(a). There was no need to search further. With the discovery that the CFCs and HCFCs depleted stratospheric ozone, the entire lower-left (chlorine-containing) portion of the triangle was eliminated (Figure 5(b)). For the one-carbon molecules, this left only R-32 (difluoromethane) with a boiling point suitable for typical refrigeration and air-conditioning applications; but R-32 is flammable. The industry turned to the two-carbon HFCs, and settled primarily on R-134a and R-125, plus blends (Figure 5(c)).

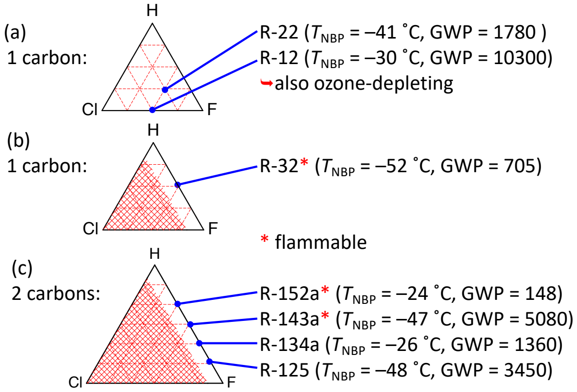

(a)
1 carbon:
H
Cl
F
R-22 (T_{NBP} = –41 ˚C, GWP = 1780)
R-12 (T_{NBP} = –30 ˚C, GWP = 10300)
also ozone-depleting
(b)
1 carbon:
H
Cl
F
R-32* (T_{NBP} = –52 ˚C, GWP = 705)
* flammable
(c)
2 carbons:
H
Cl
F
R-152a* (T_{NBP} = –24 ˚C, GWP = 148)
R-143a* (T_{NBP} = –47 ˚C, GWP = 5080)
R-134a (T_{NBP} = –26 ˚C, GWP = 1360)
R-125 (T_{NBP} = –48 ˚C, GWP = 3450)

Figure 5. Depiction of constraints on CFCs, HCFCs, and HFCs on triangular composition diagrams; (a) one-carbon compounds with no constraints: R-12 and R-22 are widely used; (b) one-carbon compounds with ODP constraint: only R-32 remains; (c) two-carbon compounds with ODP constraint: several HFCs are candidates for A/C systems.

With the additional constraint, imposed by the Kigali Amendment, of reducing the GWP of refrigerants it seems unlikely that even the GWP_{100} = 1360 of R-134a will be an acceptable long-term refrigerant, except perhaps in niche applications. This would leave only R-32 and R-152a remaining among the HFCs. The "A2" rating for R-152a represents a greater degree of flammability than many in the industry are comfortable with. R-32 is less flammable (rating of "A2L"), but it has a moderately high GWP_{100}, which may limit its long-term viability as a candidate. Once again, the search was on. But, even as the HFCs were being commercialized in the 1990s and before GWP was an explicit concern, the search for alternative fluids continued, and a wide variety of chemical classes were considered (see, for example, Bivens and Minor[57]). Concerns over global warming, and, specifically the E.U. F-gas regulations[1], reignited the search, and ultimately, the hydrofluoroolefins (HFOs) were deemed to have the best combination of characteristics.

The presence of a carbon-carbon double bond defines the chemical class of "olefin," and it is the reactivity of the double bond to the atmospheric hydroxyl radical that results in the very low atmospheric lifetimes and GWP values of the HFOs. The reactivity of the olefins, both in the atmosphere and in refrigeration equipment, is discussed by Kujak and Sorenson.[58] The most-common HFC refrigerants (R-134a, R-125, R-32, etc.) are based on one or two carbon atoms. The two-carbon HFOs, on the other hand, have normal boiling point temperatures that are too low for most refrigeration applications, as shown in Figure 6(a). Thus, most of the development efforts are focused on the three-carbon HFOs, and here there are a number of fluids with –40 ˚C < T_{NBP} < –20 ˚C; see Figure 6(b). But a three-carbon HFO is a more complicated molecule than a one- or two-carbon HFC, and this results in a higher vapor molar heat capacity and a vapor dome that is "skewed" to the right, as discussed by McLinden.[10]

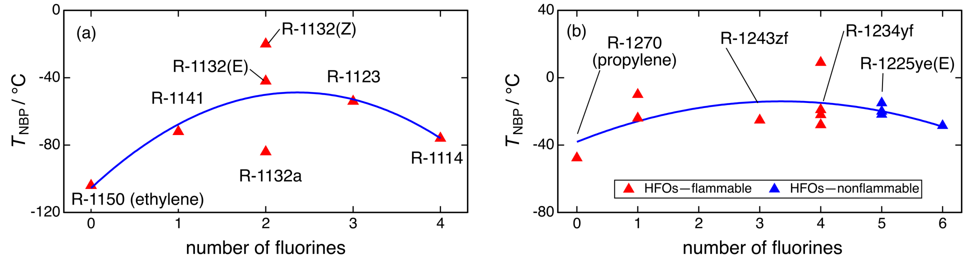


Figure 6. Variation of normal boiling point temperature T_{NBP} with the number of fluorines on the molecule; (a) two-carbon HFOs; (b) three-carbon HFOs. The curves indicate the general trends in T_{NBP}. Flammability is indicated with a color code.

Flammability versus GWP. The best-known HFO is R-1234yf, and this fluid is "marginally flammable" with an ASHRAE Standard 34 flammability classification of "2L". There is often the perception that some degree of flammability is inevitable with the HFOs. Figure 6 indicates flammability and shows that this is not true, but that, compared to the HFCs, additional fluorines are needed with the HFOs to offset the reactivity of the carbon-carbon double bond. For example, the four fluorines of R-134a are sufficient to yield a nonflammable fluid, while the four fluorines on R-1234yf are not; the three-carbon HFOs require five fluorines for nonflammability (*e.g.*, R-1225ye(E)). For the two-carbon HFOs, even the fully fluorinated R-1114 is flammable.

Among the HFCs, on the other hand, a tradeoff between flammability and GWP is somewhat inevitable. The C–F bond in the HFCs absorbs infrared radiation, and this is the root of the radiative forcing (*i.e.*, high GWP) seen with the HFCs. As the number of fluorines on the molecule increases the infrared absorption increases, but also the number of hydrogens must decrease, and it is the reaction of hydrogen with atmospheric hydroxyl that degrades HFCs in the atmosphere. This combined effect results in GWP_{100} values increasing from <1 for R-170 (ethane)[59] to 11 100 for R-116.[12] The HFOs also have C–F bonds and also absorb infrared radiation, but their much, much shorter atmospheric lifetimes result in very low GWP values. Thus, the HFOs can sidestep the flammability versus GWP tradeoff—except that the flammability versus fluorine-number relationship discussed above means that there are a smaller percentage of possible HFOs that are nonflammable.

How Reliable Was the Screening? Did We Miss Promising Fluids? The reliability and completeness of the screening could have been compromised by a number of factors. First, we considered that the PubChem database was a complete listing of the small molecules that we were interested in. While there is no guarantee that this is the case, we note that of the 31 possible three-carbon HFO isomers, 30 were listed in PubChem; it is unlikely that the missing molecule would possess significantly different properties.

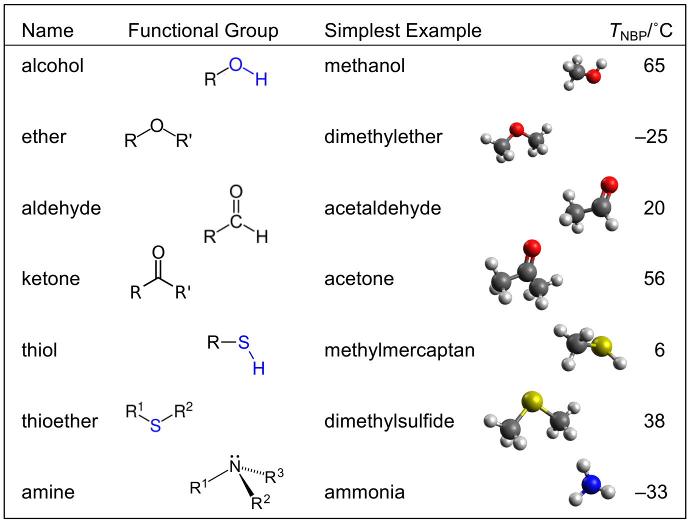
Did the restricted list of eight elements and maximum molecular size of 18 atoms exclude viable candidates? To explore this, we also carried out a search in the DIPPR database[60] of 2000 industrial chemicals based solely on 300 K ≤ T_{crit} ≤ 400 K (see McLinden *et al.*[7] for details). That search revealed 33 fluids not listed in Table 3, but these included the high-GWP CFCs and HCFCs (the DIPPR database does not tabulate GWP values) and highly toxic fluids, such as HCl, HBr, and H_2S. Eight fluids included elements not in our restricted set, but these included included radon (a radioactive gas); arsine (AsH_3), which is highly toxic and pyrophoric (ignites spontaneously on contact with air); phosphine (PF_5), which is toxic; and two silanes—SiH_3Cl and CH_3SiH_3, which are both flammable, with SiH_3Cl also being toxic. None of the molecules identified in the DIPPR search had more than 18 atoms. In other words, none of these fluids brought into question our decision to limit our search to the eight elements (C, H, F, Cl, Br, O, N, and S), although, as noted above, there is renewed interest in CF_3I. In the end, the DIPPR database did not expand the list of candidate fluids obtained from our screenings of the PubChem database.

We searched on specific ranges for the properties, and given that most of the properties were estimated from molecular structure, what was the effect of uncertainties in those estimates? The estimated uncertainty in the critical temperature was 16.5 K, but the range of T_{crit} in the screening was expanded to accommodate this uncertainty. The uncertainty in the estimate of GWP_{100} was a factor of three, and while this might seem large, it was unlikely to exclude any viable candidates—even if a candidate with an estimated GWP_{100} = 10, say, was actually 3 or 30, it would still be considered "very low GWP."

Finally, there is the possibility of incorrect data in the literature. We noted above that R-1132a was originally excluded from the original list of 27 fluids[9] based on an indication of toxicity from a reliable source.[40] Based on newer data, it was later added to the list. There is no way to avoid this situation except to periodically check for updated data.

Other fluids? But why are there so few other fluids? In chemistry, molecules are categorized by the functional group(s) they possess. The HFCs belong to the fluorinated alkane category. The alkanes comprise one or more carbon atoms with hydrogens attached to each carbon as needed to satisfy the requirement that each carbon atom be bonded to four neighbors. In the fluorinated alkanes, some of the hydrogens are simply replaced by fluorine. The alkenes (also known as olefins) are a variation on the alkanes with one or more of the carbon-carbon bonds replaced with a double bond, with each double bond "counting" as two neighbors.

Many other functional groups are well known, and some of these are shown in Figure 7. These are, for the most part, not suitable as refrigerants in moderate-to-high pressure systems because the normal boiling points for even the simplest examples are too high. The simplest alcohol, for example, is methanol with T_{NBP} = 65 ˚C. Acetone is the simplest ketone with T_{NBP} = 56 ˚C. Substitution of hydrogen with other atoms generally increases the boiling point. Functional groups with T_{NBP} low enough to serve as the basis for moderate-to-high-pressure refrigerants include the ethers (C-O-C) and the amines (comprising a central nitrogen with three attached groups). All such compounds were included in the database searches, but apart from dimethyl ether, ammonia, and a very few others, they were rejected because of high GWP_{100}, high T_{crit}, or other factors.



Name	Functional Group	Simplest Example	T_{NBP}/˚C
alcohol		methanol	65
ether		dimethylether	–25
aldehyde		acetaldehyde	20
ketone		acetone	56
thiol		methylmercaptan	6
thioether		dimethylsulfide	38
amine		ammonia	–33

Figure 7. Selected functional groups containing oxygen, sulfur, or nitrogen, with the simplest example of each also shown; R, R', R^1, R^2, R^3 indicate attached hydrogens or alkyl groups.

Conclusions

In this paper we have summarized the results of a thorough and systematic examination of possible low-GWP replacements for the HFC refrigerants currently used in small air-conditioning systems. The significant elements of this effort were (1) an optimization of fundamental thermodynamic parameters for a refrigerant; this approach was not restricted to known fluids, but encompassed all possible fluids allowed by thermodynamics (an approach we dubbed the "exploration of thermodynamic space"); (2) a search in an exhaustive chemical database for fluids having the identified thermodynamic parameters as well as acceptable chemical stability, toxicity, and GWP; and (3) simulation of the candidate fluids and a final screening based on the COP and capacity in the vapor-compression cycle. The identified list of 29 candidate refrigerants includes well-known fluids as well as fluids being actively investigated by industry; several novel fluids were also identified. But none of these are a direct, nonflammable, low-GWP replacement for the R-410A currently used in the majority of small air-conditioning systems.

The present study was focused on the medium-to-high-pressure refrigerants used in most small systems. Different equipment types may yield a different list of candidates. In particular, chillers with centrifugal compressors (which typically employ low-pressure refrigerants) were not considered, nor were high-temperature heat pumps or very-low-temperature systems. Application of the present methodology to such systems would be worthwhile.

Calm[61] introduced the concept of "refrigerant generations," and he characterizes the current, fourth generation of refrigerants as driven by "attention to global warming." Later, Calm[44] posited that, in view of limited options, a fifth generation, which would "re-examine previously discarded candidates", may be forthcoming as the full range of tradeoffs are considered. But the results of McLinden *et al.*[9] indicate (as summarized here) that there are no fundamentally new classes of chemicals available for use in vapor-compression refrigeration systems. While the "re-examination" discussed by Calm[44] may well occur, the situation may be to return (recycle?) to earlier generations, such as the ammonia, CO_2, and hydrocarbons of the first generation of "whatever works." Let us hope that we are not forced back to a "generation zero" of harvesting ice from frozen lakes.

Acknowledgements

Much of the work described in this paper was carried out in a project supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy under contract no. DE-EE002057. The work on the refrigerant blends was supported by the U.S. Department of Defense, Strategic Environmental Research and Development Program (SERDP), project WP-2740.

Author Information

M.O. McLinden, I.H. Bell and A.F. Kazakov : National Institute of Standards and Technology, 325 Broadway, Boulder, Colorado 80305, USA.

P.A. Domanski and R. Brignoli*: National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899, USA (*present address: VIS Engineering, Padova, Italy)

J.S. Brown: Catholic University of America, 620 Michigan Avenue, NE, Washington, DC 20064, USA

Dr. Mark O. McLinden was the 2018 winner of the IOR J&E Hall Gold Medal for innovation. He joined what was then the National Bureau of Standards in 1984, and his current research in the NIST Applied Chemicals and Materials Division focuses on highly accurate measurements of fluid properties over wide ranges of temperature and pressure and the design and fabrication of instruments for such measurements. Throughout his career, Dr. McLinden has researched "new" refrigerants; in the 1990s replacements for the ozone-depleting CFC and HCFC refrigerants were the focus, and more recently, his attention has turned to fluids having low global warming potential (GWP). He is the author or coauthor of more than 100 peer-reviewed publications. Dr. McLinden holds a B.S. from the University of Missouri and M.S. and Ph.D. degrees from the University of Wisconsin, all in chemical engineering.

References

1. Regulation (EU) No 517/2014 of the European Parliament and of the Council of 16 April 2014 on fluorinated greenhouse gases and repealing Regulation (EC) No 842/2006. 2014.
2. U.S. Environmental Protection Agency Significant New Alternatives Policy (SNAP): SNAP Regulations. https://www.epa.gov/snap/snap-regulations#notices (accessed December 7, 2018).
3. Kazakov, A.; McLinden, M. O.; Frenkel, M., Computational design of new refrigerant fluids based on environmental, safety, and thermodynamic characteristics. *Ind. Eng. Chem. Res.* **2012,** *51*, 12537-12548.
4. McLinden, M. O.; Domanski, P. A.; Kazakov, A.; Heo, J.; Brown, J. S. Possibilities, limits, and tradeoffs for refrigerants in the vapor compression cycle, 2012 ASHRAE/NIST Refrigerants Conference, Gaithersburg, MD, ASHRAE, Inc.: Gaithersburg, MD, 2012.
5. Domanski, P. A.; Brown, J. S.; Heo, J.; Wojtusiak, J.; McLinden, M. O., A thermodynamic analysis of refrigerants: Performance limits of the vapor compression cycle. *Int. J. Refrig.* **2014,** *38*, 71-79.
6. McLinden, M. O.; Kazakov, A. F.; Brown, J. S.; Domanski, P. A., A thermodynamic analysis of refrigerants: Possibilities and tradeoffs for Low-GWP refrigerants. *Int. J. Refrig.* **2014,** *38*, 80-92.
7. McLinden, M. O.; Brown, J. S.; Kazakov, A. F.; Domanski, P. A. In *Hitting the bounds of chemistry: Limits and tradeoffs for low-GWP refrigerants.*, 24th International Congress of Refrigeration, Yokohama, Japan, International Institute of Refrigeration: Yokohama, Japan, 2015.
8. Domanski, P. A.; Brignoli, R.; Brown, J. S.; Kazakov, A. F.; McLinden, M. O., Low-GWP refrigerants for medium and high-pressure applications. *Int. J. Refrig.* **2017,** *84*, 198-209.
9. McLinden, M. O.; Brown, J. S.; Brignoli, R.; Kazakov, A. F.; Domanski, P. A., Limited options for low-global-warming-potential refrigerants. *Nat. Comm.* **2017,** *8*, 14476.
10. McLinden, M. O., Thermodynamics of the new refrigerants. In *25th International Congress of Refrigeration*, International Institute of Refrigeration: Montreal, Quebec, 2019.
11. Bell, I. H.; Domanski, P. A.; McLinden, M. O.; Linteris, G. T., The hunt for nonflammable refrigerant blends to replace R-134a. *Int. J. Refrig.* **2019,** *104*, 484-495.
12. WMO (World Meteorological Organization), *Scientific Assessment of Ozone Depletion: 2018, Global Ozone Research and Monitoring Project–Report No. 58*. WMO: Geneva, Switzerland, 2018.
13. Velders, G. J. M.; Fahey, D. W.; Daniel, J. S.; McFarland, M.; Andersen, S. O., The large contribution of projected HFC emissions to future climate forcing. *Proc. Natl. Acad. Sci.* **2009,** *106*, 10949-10954.
14. Velders, G. J. M.; Fahey, D. W.; Daniel, J. S.; Andersen, S. O.; McFarland, M., Future atmospheric abundances and climate forcings from scenarios of global and regional hydrofluorocarbon (HFC) emissions. *Atmos. Environ.* **2015,** *123*, 200-209.
15. Velders, G. J. M., unpublished analysis based on Velders, *et al.* (2015). **2016**.
16. Houghton, J. T.; Jenkins, G. J.; Ephraums, J. J., Eds., *Climate Change: The IPCC Scientific Assessment*. Cambridge University Press: Cambridge, England, 1990.
17. Lemmon, E. W.; Bell, I. H.; Huber, M. L.; McLinden, M. O. *NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties—REFPROP, Version 10.0*, National Institute of Standards and Technology, Gaithersburg, MD: 2018.
18. Velders, G. J. M.; Andersen, S. O.; Daniel, J. S.; Fahey, D. W.; McFarland, M., The importance of the Montreal Protocol in protecting climate. *Proc. Natl. Acad. Sci. USA* **2007,** *104*, 4814–4819.
19. IPCC *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*; IPCC, Geneva Switzerland: 2014.
20. McLinden, M. O.; Didion, D. A., CFCs: Quest for Alternatives. *ASHRAE J.* **1987,** *29* (12), 32-42.

McLinden, Mark O.; Kazakov, Andrei F.; Brown, J; Brignoli, Riccardo; Bell, Ian; Domanski, Piotr A. "Options for Low-GWP Refrigerants."
Paper presented at Institute of Refrigeration, London, UK. November 06, 2019 - November 06, 2019.

21. Calm, J. M.; Didion, D. A. Trade-offs in refrigerant selection: Past, present and future, ASHRAE/NIST Refrigerants Conference: Refrigerants for the 21st Century, Gaithersburg, MD USA, October 6-7, Gaithersburg, MD USA, October 6-7, 1997; pp 6-19.

22. Kujak, S.; Schultz, K., Insights into the next generation HVAC&R refrigerant future. *Sci. Tech. Built Environ.* **2016,** *22* (8), 1226-1237.

23. ASHRAE *ANSI/ASHRAE Standard 15-2019 Safety Standard for Refrigeration Systems*; American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA: 2019.

24. Huber, M. L.; Ely, J. F., A predictive extended corresponding states model for pure and mixed refrigerants including an equation of state for R134a. *Int. J. Refrig.* **1994,** *17*, 18-31.

25. Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H., PubChem substance and compound databases. *Nucleic Acids Res.* **2016,** *44*, D1202-D1213.

26. McLinden, M. O., Optimum refrigerants for non-ideal cycles: An analysis employing corresponding states. In *USNC/IIR–Purdue Refrigeration Conference and ASHRAE-Purdue CFC Conference*, W. Lafayette, IN, July 17-20, 1990; pp 69-79.

27. Midgley, T., From the periodic table to production. *Ind. Eng. Chem.* **1937,** *29* (1), 241-244.

28. Kazakov, A.; Muzny, C. D.; Diky, V.; Chirico, R. D.; Frenkel, M., Predictive correlations based on large experimental datasets: Critical constants for pure compounds. *Fluid Phase Equilib.* **2010,** *298*, 131-142.

29. Carande, W. H.; Kazakov, A.; Kroenlein, K., Comparison of machine learning algorithms for the prediction of critical values and acentric factors for pure compounds. *251st ACS National Meeting, San Diego, March 13-17*, 2016.

30. European Environment Agency, Regulation (EU) No 517/2014 of the European Parliament and of the Council of 16 April 2014 on fluorinated greenhouse gases and repealing Regulation (EC) No 842/2006. 2014.

31. Brummond, K. M., Allene chemistry. The changing landscape of C_{sp}. *Beilstein J. Org. Chem.* **2011,** *7*, 394-395.

32. Middleton, W. J.; Sharkley, W. H., Fluoroacetylene. *J. Am. Chem. Soc.* **1959,** *81*, 803-804.

33. Lindley, A. A.; Noakes, T. J. Consideration of hydrofluoroolefins (HFOs) as potential candidate medical propellants.http://www.mexichemfluor.com/downloads_3/2010%20MF%20HFOs%20as%20candidate%20medical%20propellants%20V2%20April%202010.pdf (accessed February 14, 2013).

34. Fukushima, M.; Hashimoto, M., Next generation low-GWP refrigerants AMOLEA. *Res. Reports Ashai Glass Co. Ltd.* **2015,** *65*, 55-60 (downloaded 22 April 2016 from http://www.agc.com/rd/library/2015/65-10.pdf).

35. Luo, Y. R., *Comprehensive Handbook of Chemical Bond Energies*. CRC Press: Boca Raton, FL, 2007.

36. Lemmon, E. W.; Huber, M. L.; McLinden, M. O. *NIST Standard Reference Database 23, NIST Reference Fluid Thermodynamic and Transport Properties--REFPROP, version 9.1*, 9.1; Standard Reference Data Program, National Institute of Standards and Technology: Gaithersburg, MD, 2013.

37. Brown, J. S.; Domanski, P. A.; Lemmon, E. W. *NIST Standard Reference Database 49, CYCLE_D: NIST Vapor Compression Cycle Design Program, version 5.0*; Standard Reference Data Program, National Institute of Standards and Technology: Gaithersburg, MD, 2012.

38. Tanaka, T.; Okamoto, H.; Ueno, K.; Irisawa, J.; Otsuka, T.; Nogami, T.; Dobashi, R. Development of a new low-GWP refrigerant composed of HFO-1123 (trifluoroethylene), AIChE Annual Meeting, Atlanta, GA, American Institute of Chemical Engineers: Atlanta, GA, 2014.

39. Hashimoto, M.; Otsuka, T.; Fukushima, M.; Okamoto, H.; Hayamizu, H.; Ueno, K.; Akasaka, R., Development of New Low-GWP Refrigerants: Refrigerant Mixtures Including HFO-1123. *Sci. Tech. Built. Environ.* **2019,** *25*, 776-783.

40. National Institute for Occupational Safety and Health, *NIOSH Pocket Guide to Chemical Hazards*. U.S. Department of Health and Human Services: 2007.

41. ASHRAE, Addendum f to ANSI/ASHRAE Standard 34-2016: Designation and Safety Classification of Refrigerants. American Society of Heating, Refrigerating and Air-Conditioning Engineers: 2017.

42. McGowan, M. K., Progress report on alternative refrigerants. *ASHRAE J.* **2019,** *61* (2), 38-41.

43. ASHRAE, Addendum t to Standard 34-2019, Designation and Safety Classification of Refrigerants. American Society of Heating, Refrigerating and Air-Conditioning Engineers: 2019.

44. Calm, J. M. Refrigerant transitions … Again, 2012 ASHRAE/NIST Refrigerants Conference, Gaithersburg, MD, ASHRAE, Inc.: Gaithersburg, MD, 2012.

45. Nimitz, J.; Lankford, L. Refrigerants containing fluoroiodocarbons (FICs), 1994 International Refrigeration Conference, W. Lafayette, IN, July 19-22, W. Lafayette, IN, July 19-22, 1994; pp 255-260.

46. Brown, J. S., Introduction to hydrofluoro-olefin alternatives for high global warming potential hydrofluorocarbon refrigerants. *HVAC&R Research* **2013,** *19*, 693-704.

47. Brown, J. S.; Brignoli, R.; Domanski, P. A. *CYCLE_D-HX: NIST Vapor Compression Cycle Model Accounting for Refrigerant Thermodynamic and Transport Properties, Version 1.0.*; National Institute of Standards and Technology: Gaithersburg, MD, 2017.

48. Brignoli, R.; Brown, J.S.; Skye, H.M.; Domanski, P.A. *Refrigerant Performance Evaluation Including Effects of Transport Properties and Optimized Heat Exchangers*, Int. J. Refrig. 2017, 80, 52-65

49. Brown, J. S.; Yana-Motta, S. F.; Domanski, P. A., Comparative analysis of an automotive air conditioning system operating with CO2 and R134a. *Int. J. Refrig.* **2002,** *25*, 19-32.

50. Schultz, K. Replacements for R410A with GWPs less than 300, 25th International Congress of Refrigeration, Montreal, International Institute of Refrigeration: Montreal, 2019; pp 531-538.

51. Spletzer, S.; Medina, J.; Hughes, J. Examining low global warming potential hydrofluoroolefin-hydrofluorocarbon blends as alternatives to R-404A, 25th International Congress of Refrigeration, Montreal, International Institute of Refrigeration: Montreal, 2019; pp 1086-1093.

52. ASHRAE *ANSI/ASHRAE Standard 34-2019 Designation and Safety Classification of Refrigerants*; American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, GA: 2019.

53. Linteris, G. T.; Bell, I.; McLinden, M. O., An empirical model for refrigerant flammability based on molecular structure and thermodynamics. *Int. J. Refrig.* **2019,** *104*, 144-150.

54. ASTM International *ASTM E-681, Standard Test Method for Concentration Limits of Flammability of Chemicals (Vapors and Gases).* ASTM International: West Conshohocken, PA, 2009.

55. Midgley, T.; Henne, A. L., Organic fluorides as refrigerants. *Ind. Eng. Chem.* **1930,** *22* (5), 542-545.

56. Downing, R. C., *Fluorocarbon refrigerants handbook*. Prentice-Hall, Inc.: Englewood Cliffs, New Jersey, 1988.

57. Bivens, D. B.; Minor, B. H., Fluoroethers and other next generation fluids. *Int. J. Refrig.* **1998,** *21*, 567-576.

58. Kujak, S.; Sorenson, E., Haloolefins Refrigerants. *ASHRAE J.* **2018,** *60* (6), 28-35.

59. Hodnebrog, Ø. i.; Dalsøren, S. B.; Myhre, G., Lifetimes, direct and indirect radiative forcing, and global warming potentials of ethane ($C_2H_{6)}$, propane (C_3H_8), and butane (C_4H_{10}). *Atmos. Sci. Lett.* **2018,** *19*, e804.

60. DIPPR, *Evaluated Standard Thermophysical Property Values, 801*. Design Institute for Physical Properties, American Institute of Chemical Engineers: 2011.

61. Calm, J. M., The next generation of refrigerants—Historical review, considerations, and outlook. *Int. J. Refrig.* **2008,** *31*, 1123-1133.

McLinden, Mark O.; Kazakov, Andrei F.; Brown, J; Brignoli, Riccardo; Bell, Ian; Domanski, Piotr A. "Options for Low-GWP Refrigerants."
Paper presented at Institute of Refrigeration, London, UK. November 06, 2019 - November 06, 2019.

Automatic Recognition of Advanced Persistent Threat Tactics for Enterprise Security

Anonymous Author(s)

ABSTRACT

Advanced Persistent Threats (APT) has become the concern of many enterprise networks. APT can remain undetected for a long time span and lead to undesirable consequences such as stealing of sensitive data, broken workflow, and so on. To achieve the attack goal, attackers usually leverage specific tactics that utilize a variety of techniques. This paper explores the recognition of APT tactics through synthesized analysis and correlation of data from various sources. We propose a framework for detecting the APT tactics and discuss the application of machine learning techniques in this problem. Our framework can be used by the security analysts for effective detection of APT attacks. The evaluation of our approach shows that it can detect APT tactics with high accuracy and low false positive rate. Therefore, it can be used for tactic-centric APT detection and effective implementation of cyber security response operations.

KEYWORDS

Advanced Persistent Threat, Attack Tactics, Machine learning.

1 INTRODUCTION

Cyber attacks against organizations, including Advanced and Persistent Threats (APT), usually employ certain attack tactics. It is common that some attack tactics are used repeatedly in different APT attacks. Identifying these tactics may help understand attackers' potential intent, objectives and strategies, and even help with identification of specific attacker(s) or attack communities. However, the adversary groups often constantly update their tools and tactics to make detection and analysis more difficult. For instance, the DragonOK group, a well-known adversary group, has been evolving their tactics in targeted APT attacks across Asia Pacific and Japan. They began to use multiple new variants of malware "FormerFirstRAT" along with malware "IsSpace" and "Tidepool" in their 2017 tactics [1].

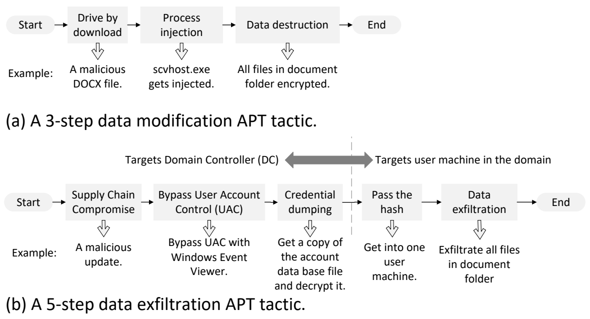
The emergence of new APT tactics has introduced daunting challenges to APT detection. The attackers can leverage a carefully designed combination of various *APT techniques* (e.g., spear phishing, drive by download, buffer overflow, pass the hash) to strategically achieve a goal. Hence, the detection of individual APT techniques is no longer adequate to identify the attacker's intents, objectives and strategies. That is, individual APT techniques cannot tell the "whole story" of the attacks. This inability has put real-world Cyber Security Operation Centers (CSOCs) into a highly undesired dilemma: (a) on one hand, without knowing the "whole story", CSOCs are more likely to take ineffective intrusion response actions; (b) on the other hand, correlating the detected individual APT techniques to generate the "whole story" requires significant amount of time and manual efforts.

A number of research works have explored the problem of detecting APT attacks from different angles. 1) Since APT may remain stealthy for a long time span, capturing all stages of its life cycle is not an easy task. Hence, some research works propose to detect a specific technique that is used in a stage of APT. For example, [11, 17] detect the network connections during the stage of malware command and control (C&C) communication. [17] applies supervised learning towards the web proxy logs to identify and prioritize the enterprise malicious activities, while [11] proposes unsupervised detection of C&C communications based on the web request graphs. [21] detects the APT malware infection by analyzing malicious DNS and network traffic. 2) Some other works aim to detect APT attack as a whole. For example, [7, 20] use classification models for APT detection; [9, 10, 19] reconstruct the attack by combining past security events. HOLMES [16] also leverages the correlation between suspicious information flows for APT detection. 3) Another angle of detecting APT attacks is through provenance tracking. A provenance tracking system captures the causality relationship between system objects such as processes and files. Security analysts can find the root cause of attacks by tracking system object dependencies generated by the provenance data. Because most existing provenance tracking techniques are at low system level and usually suffer from dependence explosion problem, some research works propose to partition execution to units [12, 14]. [13] further proposes to leverage the annotated application specific high level task structures to partition execution. 4) Mining logs is another technique that is commonly used for APT attack detection, although the purposes and approaches of mining may be different. [18] proposes an automated multi-stage intrusion analysis system that is based on mining various logs. The proposed system discovers the "attack communities" from the weighted graphs that are built from multiple logs. [8] proposes to use a deep neural network model, DeepLog, to detect anomaly log sequences. It models a system log as a natural language sequence, learns the log patterns from normal system execution, and reports anomaly when the log patterns deviate from the trained model.

Although extensive researches have been conducted towards detecting APT, few of them emphasize the detection of commonly used APT tactics. For approaches that focus on specific steps or techniques such as C&C communication, or that focus on log mining, the detected anomaly may or may not be relevant to APT. Other approaches, that focus on provenance tracking, event correlation or clustering, are not tactic-centric.

However, identifying tactics is essential to reveal attackers' potential objectives and strategies, and may even help to identify specific adversary groups. Therefore, this paper seeks to propose a framework that can detect APT tactics with high accuracy and general applicability. If successful, the framework is able to: (a) identify the APT techniques that are not easily detected with traditional approaches; (b) match the APT techniques to the tactics they belong to with the help of system object dependencies; (c) make the APT

Zou, Qingtian; Singhal, Anoop; Sun, Xiaoyan; Liu, Peng. "Automatic Recognition of Advanced Persistent Threat Tactics for Enterprise Security." Paper presented at 6th ACM International Workshop on Security and Privacy Analytics 2020, New Orleans, LA, US. March 16, 2020 - March 18, 2020.


Start
Drive by download
Process injection
Data destruction
End
Example:
A malicious DOCX file.
scvhost.exe gets injected.
All files in document folder encrypted.
Targets Domain Controller (DC)
Targets user machine in the domain
Start
Supply Chain Compromise
Bypass User Account Control (UAC)
Credential dumping
Pass the hash
Data exfiltration
End
Example:
A malicious update.
Bypass UAC with Windows Event Viewer.
Get a copy of the account data base file and decrypt it.
Get into one user machine.
Exfiltrate all files in document folder

(a) A 3-step data modification APT tactic.

(b) A 5-step data exfiltration APT tactic.

Figure 1: Two example APT tactics.

tactic and APT technique identifiers extensible and adaptive to new tactics and techniques.

The significance of this research is three-folds: 1) it is, to the best of our knowledge, the first framework which could simultaneously achieve accuracy and general applicability in detecting multiple APT tactics; 2) implementations of the framework could help CSOCs and analysts identify the attacker's intents, objectives and strategies, and provide a "whole story" of the attacks; 3) the automated APT tactic identifier could significantly reduce the manual efforts involved in detecting APT techniques.

The remaining of this paper is structured as follows. Section 2 will discuss the APT tactic with two examples, and also the differences from attack graph, which is a well-known graphical method in cyber security. Section 3 will discuss our proposed framework. Section 4 will briefly introduce our design and implementation. In Section 5, a simple five-step APT tactic is presented and used to demonstrate how our framework works. Section 6 presents our evaluation experiments results. Section 7 is the conclusion.

2 PRELIMINARY

The MITRE adversarial tactics and techniques knowledge repository [5] provides a comprehensive review of the real-world adversarial tactics and techniques. In this work, we adopt definitions for tactic and technique different from those in [5]. We define:

- *APT technique*: A specific implementation such as a hacking tool, attack script, and/or malware payload.
- *System object dependency*: The relationships among system objects, such as processes, files and user accounts.
- *APT tactic*: A sequence of APT techniques chained by system object dependencies.

Therefore, APT tactics represent attackers' strategies; and APT techniques represent the specific steps that attackers take to implement the strategies.

In this paper, we present an APT tactic as a connected graph showing the chain of techniques in a multi-step cyber attack. APT techniques are basic building blocks of an APT tactic. All the APT tactic presented in this paper are crafted based on our observation on different attack scenarios. Figure 1 presents two example APT tactics as directed graphs that consist of multiple APT techniques.

Each technique in a tactic has its post-conditions and pre-requisites. Post-conditions are the results of the technique, such as malicious processes being created, files being accessed and user account being modified. Pre-requisites describe the requirements for the technique to be matched into tactic.

In Figure 1(a) and Table 1, an APT tactic about data destruction is presented. This tactic starts from drive by download. In this technique, the download is requested by the user, but the downloaded item include functionalities that user does not expect. The user downloads the program and executes it, without knowing that the program has a trojan built in. The trojan allows a remote attacker to connect and execute malicious commands on the victim computer. With the remote access, the attacker can then escalate the privilege to the system level by process injection. Afterwards, the attacker can destruct all documents by deleting or encrypting them.

In Figure 1(b) and Table 2, an APT tactic about data exfiltration is presented. The first three techniques are launched against Windows Domain Controller (DC), whereas the last two are against another user machine in the domain. This tactic starts with supply chain compromise. For privilege escalation, the attacker bypasses the User Account Control (UAC). UAC prompts user for confirmation when a process requests for system-level privileges. By bypassing UAC, the attacker can escalate privileges without being noticed. After that, the attacker dumps users' credentials such as password hashes. These credentials can be used to launch pass the hash attacks to access other machines. In the end, the attacker downloads sensitive files from target machines.

APT tactics are fundamentally different from attack graphs. Some important differences include: (a) attack graphs represent the causality relationship between vulnerabilities and exploits, whereas APT tactics represent the strategies, techniques and procedures used by attackers; (b) attack graphs show all the possible attack routes from the attacker's machine to the target machine, whereas APT tactics focus on the attackers' chosen techniques and procedures, rather than the attack paths; (c) attack graphs are not being used by CSOCs on a daily basis, whereas APT tactics are frequently referred to by security analysts, though in an implicit and informal way according to our observation.

3 PROPOSED FRAMEWORK

To serve the validity of the proposed framework, we assume that:

- Each APT technique used in the APT tactic is identifiable through automated, semi-automated, or manual effort. In the worst case, the CSOC may have to resort to manual effort to identify a particular APT technique, we have no assumption on the maximum time used to identify an APT technique.
- The whole framework is kept safe from the attacker. All the input data is genuine, which means that the attackers cannot modify or delete them; and the attackers have no access to the framework itself in any way.

The framework presented above needs to address the following challenges:

- *Accurate identification of various adversary techniques.* Although we assume every APT technique is identifiable, some APT techniques, such as pass the hash, are hard to accurately identify with traditional methods like pattern matching and

Table 1: Detailed description of a 3-step APT tactic with system object dependency included.

Technique Name	**Post-condition**	**Pre-requisites**	**Description**	**Identification Method**
Drive by download	Process P_1 is created.	(None)	Initial intrusion by drive by download attack, resulting to a malicious process P_1 being created.	The downloaded item has some kind of backdoor built-in. Some backdoors are detectable using signature-based IDS.
Process injection	Process P_2 is created; Process P_3 is affected.	Process P_2 is a child process of or the same as P_1.	Privilege escalation by process injection. Process P_2 is the process which initiates the injection, and P_3 is the victim process. P_3 is usually a process running with system-level privilege.	Detectable by monitoring critical system processes.
Data destruction	Process P_4 is created; File F_1 is accessed.	Process P_4 is a child process of or the same as P_3.	Malicious process P_4 accesses file F_1 and makes it inaccessible to the user.	Detectable by monitoring disk I/O on sensitive files/directories.

Table 2: Detailed description of a 5-step APT tactic with system object dependency included.

Technique Name	**Post-condition**	**Pre-requisites**	**Description**	**Identification Method**
Supply chain compromise	Process P_1 is created.	(None)	Initial intrusion by supply chain compromise, resulting to a malicious process P_1 being created.	Some software distribution or update channels get infected and backdoor gets inserted to the products. Some backdoors are detectable using signature-based IDS.
Bypass User Account Control (UAC)	Process P_2 is created; Process P_3 is created.	Process P_2 is a child process of or the same as P_1.	The attacker bypass the Windows UAC to escalate its privilege to system level.	Many procedures of bypassing UAC needs to modify the Windows registry. Such procedures can be detected by monitoring the registry for specific key creation and modification.
Credential dumping	Process P_4 is created.	Process P_4 is a child process of or the same as P_3.	The attacker leverages escalated privilege to dump user credentials like password hashes.	In a Windows Domain, the DC stores the users' credentials as a database file, and very few processes are allowed to interact with this file. Dumping users' credentials can be detected by monitoring the disk I/O on this file and activities of those special processes.
Pass the hash	Process P_5 is created; User U_1 is impersonated.	(None)	The attacker leverages the password hashes to get into other machines in this domain.	Directly using hashes for authentication relies on certain authentication mechanism, which will leave traces in the network packets. Thus, it is detectable by monitoring the network traffic and inspecting the network packets.
Data exfiltration	Process P_6 is created; File F_1 is read.	Process P_6 is a child process of or the same as P_5.	The attacker, pretending to be user U_1, downloads file F_1 to his/her own machine.	Detectable by monitoring disk I/O on sensitive files/directories.

anomaly detection. Pattern matching suffers from low accuracy, and anomaly detection suffers from high false positive rate.

- *Correct match of the adversary techniques to the tactics they belong to.* Assuming that each technique can be accurately identified, the APT tactic matcher needs to match those techniques into tactics. For receiving inputs, the matcher needs to deal with the diversity of technique identifiers, such as different identification delay; for matching, the matcher needs to deal with multiple cases, namely (a) one attacker is using one tactic, (b) multiple attackers are using one same tactic, and (c) multiple attackers are using different tactics. The framework should address those problems.

For the first challenge, we propose to apply machine learning method for those APT techniques that cannot be accurately identified with traditional methods. Pass the hash is one of such techniques. It is difficult to identify with traditional methods because it leverages legitimate authentication mechanism. To apply machine learning method, a huge amount of data is needed to train an accurate neural network. Because we didn't find any open network log data sets for identifying pass the hash, we generate our own data set. Details are described in section 5. The evaluation of our trained neural network is presented in subsection 6.1.

For the second challenge, we design a framework as shown in Figure 2. It comprises three concurrent workflows, the data processing workflow, the tactic knowledge processing workflow, and system object dependency discovering workflow, to address the tactic matching problem.

The data processing workflow is as follows:

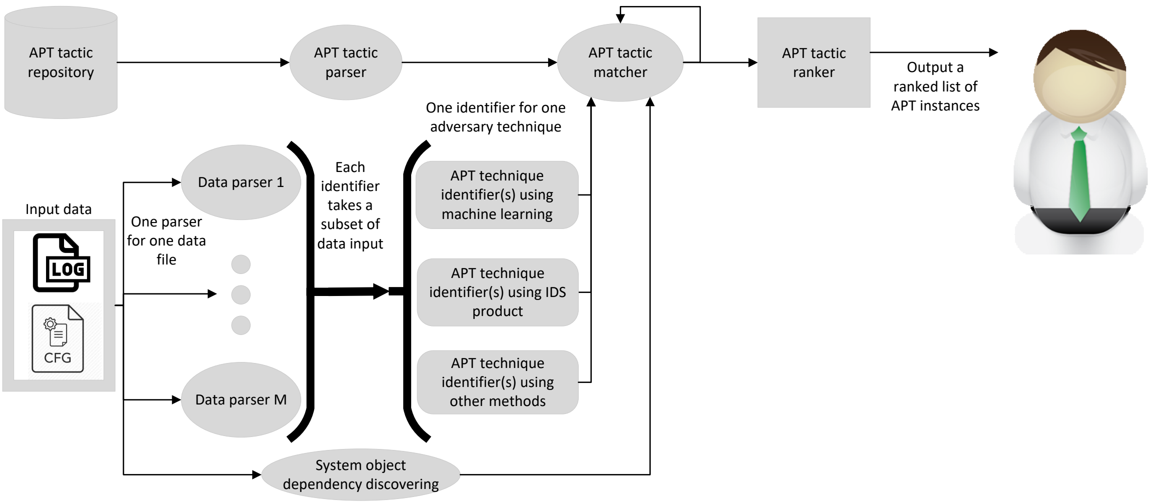

APT tactic repository
APT tactic parser
APT tactic matcher
APT tactic ranker
Output a ranked list of APT instances
One identifier for one adversary technique
Input data
LOG
CFG
One parser for one data file
Data parser 1
Data parser M
Each identifier takes a subset of data input
APT technique identifier(s) using machine learning
APT technique identifier(s) using IDS product
APT technique identifier(s) using other methods
System object dependency discovering

Figure 2: Proposed architecture.

(1) **Data Parsing.** The collected data sources are first fed to the corresponding data parsers.
(2) **Technique Identifying.** Based on parsed data, the APT technique identifiers determine whether certain adversary techniques exist or not.

The tactic knowledge processing workflow is as follows:

- **Tactic Parsing.** The previously seen APT tactics, which are stored in the APT tactic repository, are fed to the APT tactic parser.

The system object dependency discovering workflow is as follows:

- **Discovering system object dependencies.** A number of system logs, such as process and file I/O monitoring logs, are used to discover system object dependencies.

Finally, the results from above three workflows are taken as input for the following procedures:

(1) **Tactic Matching.** The APT tactic matcher uses the parsed tactics to match the identified adversary techniques. The matched tactics are stored as tactic instances, no matter it is fully matched or partially matched.
(2) **Tactic Ranking.** All APT tactic instances are ranked based on completeness of tactic matching.

4 DESIGN AND IMPLEMENTATION

The framework is implemented in a Ubuntu 18.04 virtual machine (referred to as detector). To isolate the detector from the IT system being monitored, it is setup as an HTTP server to receive file upload. On receiving files upload, the files are put to specific directories based on its type, such as APT tactic files, network log files or windows event log files. Other daemon programs, which are configured to monitor those directories, will trigger the framework to run once the directories' contents are changed. It is also possible to set the triggering to manual if the contents are changed at a high frequency. In this case, the daemon program monitoring the contents can raise alerts to notify the security analysts about the arrival of new input files. The analysts can then run the program at a preferred time point.

Data parser. On receiving system information, each file (e.g. a log file or configuration file) is assigned to the corresponding data parser for processing. Different data parsers are designed to deal with different types of input system files based on their syntaxes. Therefore, the number of data parser types is the same as the number of input system file types. The data parsers can work in parallel for faster processing speed.

APT technique identifier. The APT technique identifiers receive parsed data from data parsers and determine whether certain APT techniques are used. Each identifier is responsible for checking one technique. Based on different APT techniques, the identifiers may need data from different sources (i.e., different types of system information files). Hence, identifiers may take data from different sets of data parsers. The identifiers can be signature based, anomaly detection based, machine learning based, or other types. The output of the identifiers is data tuples made up of technique name and its post-conditions. Take the "Data exfiltration" technique in Table 2 as an example, the result output can be `("Data exfiltration", 23619, "D:\Documents\customer-list.xlsx")`, in which `23619` is the process ID (PID) on the user machine, and `"D:\Documents\customer-list.xlsx"` is the file. The malicious process with ID `23619` reads the sensitive file `"D:\Documents\customer-list.xlsx"` in this technique.

APT tactic repository. The APT tactic repository is a directory where all APT tactic files are stored. All the previously seen APT tactics are stored in the APT tactic repository in the same syntax. We use the graph description language DOT to describe these APT tactics. In APT tactic graphs, every adversary technique is defined as a box-shaped node, and directed edges denote the attack order.

Zou, Qingtian; Singhal, Anoop; Sun, Xiaoyan; Liu, Peng. "Automatic Recognition of Advanced Persistent Threat Tactics for Enterprise Security." Paper presented at 6th ACM International Workshop on Security and Privacy Analytics 2020, New Orleans, LA, US. March 16, 2020 - March 18, 2020.

Each DOT file can be easily visualized into directed graphs by tools like Graphviz [3]. In cases where other syntax (e.g. STIX [6]) are needed, a new APT tactic parser should be added correspondingly to parse tactic files in the new syntax.

Listing 1 is the DOT description of the APT tactic shown in Figure 1 (a) and Table 1. The example graph contains 3 technique nodes, 5 post-condition nodes, and 3 pre-requisite nodes. Each node definition starts with a *node ID*, and follows by node attributes such as *label* (the text to show in graph), *shape* (the node shape in graph) and *style* (the node style in graph). We use rounded boxes to denote the start and end nodes of the graph, boxes for APT techniques, triangles for post-conditions, and inverted triangles for pre-requisites. Post-conditions are presented as result system objects, and pre-requisites are presented as the relationships between system objects, such as "P1=>P2" stands for process P_2 is a child process of or the same as P_1, "P2->F1" stands for process P_2 writes file F_1, and that "F1->P3" stands for process P_3 reads file F_1.

Listing 1: DOT codes example.

```
digraph example_tactic {
    // nodes
    1[label="start",shape=box,style=rounded];
    2[label="Drive by download",shape=box];
    21[label="P1",shape=triangle];
    3[label="Process injection",shape=box];
    31[label="P2",shape=triangle];
    32[label="P3",shape=triangle];
    33[label="P1=>P2",shape=invtriangle];
    4[label="Data destruction",shape=box];
    41[label="P4",shape=triangle];
    42[label="P5",shape=triangle];
    43[label="P4=>P5",shape=invtriangle];
    44[label="P5->F1",shape=invtriangle];
    5[label="end",shape=box,style=rounded];

    // edges
    1->2->3->4->5;
    21->2;
    31->3;32->3;33->3;
    41->4;42->4;43->4;44->4;
}
```

APT tactic parser. From APT tactic DOT files, the APT tactic parser extracts information including: (a) the number of nodes in the tactics; (b) the types of these nodes; (c) the way that the nodes are connected; (d) the post-conditions and pre-requisites for each technique in this tactic. The information is used to create the initial "template" APT tactics. They are "templates" because no identified techniques have been matched into them at this time. We store the tactic templates in a multi-layer JSON-like data structure, which features pairs of name and values. The first layer presents technique node IDs in the DOT file, and the second layer presents node properties such as node name, post-conditions and pre-requisites. For example, the parsed APT tactic for the tactic in Figure 1(a) and Table 1 is presented below.

Listing 2: Parsed APT tactic example

```
{   '1': {  'name': 'start',
            'post-conditions': 'none',
            'pre-requisites': 'none'},
    '2': {  'name': 'Drive by download',
            'post-conditions': 'P1',
            'pre-requisites': 'none'},
    '3': {  'name': 'Process injection',
            'post-conditions': 'P2; P3',
            'pre-requisites': 'P1=>P2'},
    '4': {  'name': 'Data destruction',
            'post-conditions': 'P4; F1',
            'pre-requisites': 'P3=>P4'},
    '5': {  'name': 'end',
            'post-conditions': 'none',
            'pre-requisites': 'none'}}
```

System object dependency discovering. Some logs, such as process and file I/O monitoring logs, are used to discover system object dependencies, which are used for chaining APT techniques in tactic matching. For example, in Figure 3, there are three edges showing a process forks to create a child process, a process writes a file, and that a process reads a file. We use triple-element tuples to present these dependencies. The first and last element each represents a system object, such as a file or a process; and the middle element represents the operation, such as process forking or file reading/writing. Each system object element is a tuple which starts with a system object indicator that represents its type, followed by other items that vary according to the system object's type. For example, file writing in Figure 3 is presented as `(("P", 5414), "write", ("F", "D:\apt.dll"))`. The first element indicates a process with PID 5414, the second element indicates file writing operation, and the third element indicates the file `"D:\apt.dll"`. Therefore, this whole entry of system object dependency means that a process with PID 5414 writes the file `"D:\apt.dll"`.

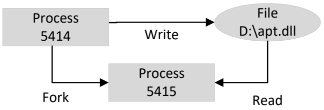


Figure 3: A simple system object dependency graph (SODG).

APT tactic matcher. The matcher takes five inputs: (a) initial "template" APT tactics from APT tactic parser; (b) APT tactic instances; (c) results from APT technique identifiers; (d) a pool of APT techniques that are identified; (e) system object dependencies. Inputs (b) and (d) are from previous run of the matcher.

Once a new APT technique is identified, the matcher will try to match it to an APT tactic if the technique is part of the tactic. Specifically, the technique can be matched to either APT tactic template(s) or partially matched APT tactic instance(s). 1) If there were partially matched APT tactic instances expecting this technique, the matcher will check the pre-requisites of the identified technique to see if they are satisfied. The pre-requisites are checked through system object dependencies. If all pre-requisites are satisfied, the new APT technique is matched into the tactic instance, and the post-conditions are confirmed in the instance. If not, the unmatched new technique is saved to the pool and the post-conditions are discarded. 2) In many times, a given APT technique may not be matched to any partially matched APT tactic instances, but can be matched to a APT template. In this case, a new APT tactic instance is created from the template, the technique is matched into this instance, and system object(s) in the instance regarding this technique is confirmed. Otherwise, the technique is put to the pool.

Whenever an APT technique is matched into an APT tactic template/instance, the matcher goes through all techniques in the pool and start another iteration of the matching process, but this time, the matcher only tries to match the technique into partially matched tactic instances, because those techniques have already been checked against templates before put into the pool. In this way, even if a technique is not matched at the time of identification, it will still be matched once all the conditions are met afterwards.

To illustrate the algorithm, we present two examples of matching based on the tactic presented in Figure 1(a) and Table 1. Firstly, consider the case where no techniques are matched into the tactic. Assuming the identifier first reports that drive by download and the corresponding process are identified. The matcher first assumes that the post-condition P_1 is the process reported, and then checks the pre-requisites. The matcher finds out that this technique has no pre-requisites, so the technique gets matched immediately. An APT tactic instance gets created from this template, the drive by download technique in it is marked as matched, and P_1 is confirmed to be the process reported in this APT tactic instance.

In the second example, consider the case where no techniques are matched into the tactic. However, technique process injection gets identified first. Similarly, the matcher assumes the system objects reported are correct, and then checks the pre-requisites against system object dependencies. The matcher finds out that it needs P_1 and P_2, but P_1 from the previous technique is not matched yet. Therefore, the matcher decides that this technique cannot be matched and put it in the pool of unmatched techniques. After drive by download is matched, during the going through of the pool, this previously unmatched technique, process injection, will be matched if it meets the pre-requisites.

In real world, defense against APT tactics generally needs to tackle three scenarios: one attacker is using one tactic; multiple attackers are using one same tactic; multiple attackers are using different tactics. Our matcher can handle all these scenarios properly. (a) If one attacker is using one tactic, the algorithm shown in Algorithm 1 can detect it by matching every step. (b) If multiple attackers are using one same tactic, they can be differentiated by system object dependencies, such as different process ID, which is presented as pre-requisites (Table 1) used in line 6 of the algorithm. (c) If multiple attackers are using different tactics, they can be differentiated by matching different techniques as shown in line 2 of the algorithm. A more complicated situation is that multiple attackers are using different APT tactics, but these tactics share some techniques in common. At the defender's side, it is unknown which attacker is using which APT tactic, so the best practice is to list all the possible combinations. In light of this, the matcher will create APT tactic instances for all combinations while preserving the system object dependencies.

Algorithm 1 APT tactic matching algorithm

```
1:  **Input**: APT tactic templates; APT tactic instances; one newly identified adversary technique $new\_at$; a pool of identified techniques.
2:  $all\_candidates \leftarrow$ same techniques as $new\_at$ found in APT tactic templates and instances.
3:  if there exists at least one technique in $all\_candidates$ then
4:      for all $candidate \in all\_candidates$ do
5:          Assuming post-conditions are right, examine pre-requisites for each $candidate$.
6:          if All pre-requisites are met for the $candidate$ then
7:              if $candidate$ is in a template then
8:                  Create a new instance from this template, and match $new\_at$ into the position of $candidate$.
9:              else if $candidate$ is in an instance then
10:                 Create a copy of $candidate$.
11:                 Match $new\_at$ into the position of the copy of $candidate$.
12:             end if
13:             In the instance where $new\_at$ is matched into, identify the next adversary technique $next\_at$.
14:             $all\_next\_step\_candidates \leftarrow$ same techniques as $next\_at$ found in the pool of not matched adversary techniques.
15:             for all $next\_step\_candidate \in all\_next\_step\_candidates$ do
16:                 Call APT tactic matching algorithm.
17:             end for
18:         end if
19:         Put $new\_at$ into the pool
20:     end for
21: else
22:     Save $new\_at$ to the pool of not matched adversary techniques.
23: end if
```

APT tactic ranker. The ranker takes the APT tactic instances from matcher as input. The ranking is based on the percentage of APT tactic instances' completeness. The more completely an APT tactic instance is matched, the higher it will appear on the list. Fully matched APT tactic instances are put on top of the list.

The ranker ranks both fully matched and partially matched APT tactic instances. Assuming that the technique identifiers can correctly identify APT techniques, the existence of partially matched APT tactic instances means that either the attacker gives up this campaign, or that the attacker just decides to wait before launching remaining techniques. Therefore, partially matched APT tactics should be kept and ranked but not discarded.

In a word, the framework's workflow can be summarized as input parsing, technique identifying, tactic matching and tactic ranking. Except input parsing phase, the other three phases generate new findings at three time points respectively, and new findings at a previous time point should lead to new findings at the next time point.

- The first time point is when a technique is identified at technique identifying phase. The new findings here are the identified techniques and their post-conditions.
- The second time point is at the end of tactic matching phase. The new findings here are are new and/or updated APT tactic instances.
- The third time point is at the end of tactic ranking phase. The new findings here are updated APT tactic instances' completeness and ranking results.

Thus, at the end of technique identification phase, if new technique(s) is identified, the tactic matcher should be automatically triggered; and at the end of tactic matching phase, if new APT tactic instance(s) is created or previous instance(s) is updated, the tactic ranker should be automatically triggered.

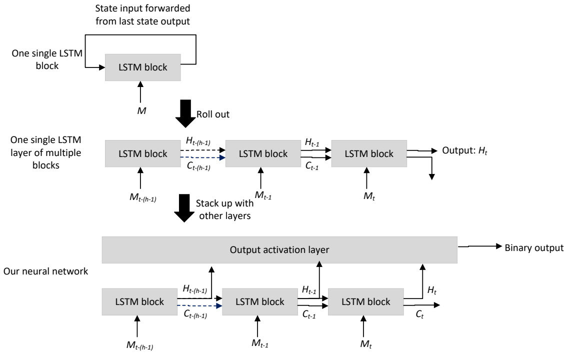


Figure 4: Our LSTM neural network.

5 CASE STUDY

The APT tactic shown in Figure 1(b) and Table 2 contains five different APT techniques. In this section, we use this tactic as a case study and describe how each APT technique can be identified, following the order they appear. We will discuss the identification of Pass the Hash in detail, as it is an example APT technique that we can apply machine learning for its identification.

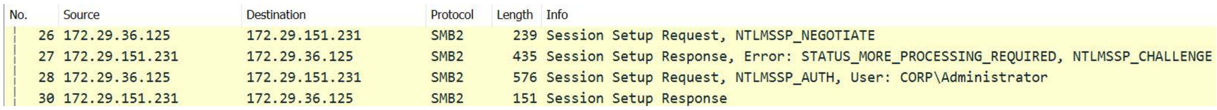
Supply Chain Compromise. This technique can be identified by scanning downloaded item with Anti-Virus (AV) products. Alternatively, network intrusion detection products like Snort [4] can also detect it when the downloaded item tries to establish communication with the attacker's machine. Windows Defender can identify this technique and leave one entry in Windows event logs. The post-condition of this step is a malicious process P_1 is created due to attackers' communication with the victim machine. As an initial intrusion step, it has no pre-requisites.

Bypassing Windows User Account Control (UAC). Bypassing Windows UAC is a common privilege escalation technique towards Windows machines. Many procedures have been discovered to bypass UAC. An extensive list of available procedures can be found in the UACMe project [2]. Some procedures rely on modifying specific, user-accessible Registry entries. Therefore, by monitoring accesses to the Registry with process monitor, especially entry creation and modification, this APT technique can be identified. The post-conditions are that processes P_2 and P_3 are created, with P_3 running at system-level privilege. P_2 may act as a middle stage. In this case, P_2 may be the actual process modifying the Registry. After that, a system service reads the modified Registry and created the process P_3 with system-level privilege. Therefore, P_2 and P_3 may not have direct relation, and the pre-requisite only has requirement on P_2.

Credential Dumping. Credential dumping is the process of obtaining account credentials, normally in the form of account names and password hashes. On a Domain Controller (DC), Active Directory (AD) service maintains all the users' account names and password hashes in the domain. They are stored in a database file locked by AD. Only some specific processes are allowed to access the contents of this file. Local Security Authority Subsystem Service (LSASS) is one of such special processes. It is given the right to read the database file because its role in a DC is to provide an interface for managing local security, domain authentication, and AD processes. With escalated privilege, attackers can craft queries, send them to LSASS, and get domain user account names and corresponding hashes. Therefore, we can monitor the LSASS process to detect this APT technique. We can either match some signatures in malicious queries or detect anomaly behaviors of LSASS. The post-condition is that process P_4 that runs with system-level privilege and interacts with LSASS is created. The pre-requisite is that P_4 is a child process of or the same as P_3, because a process needs system-level privilege to interact with LSASS.

Pass the Hash. Pass the hash is a well-known technique for lateral movement. In remote login, plain-text passwords are usually converted to hashes for authentication. Some authentication mechanisms only check if hashes are matched. Pass-the-hash technique relies on these vulnerable mechanisms to impersonate a normal user with dumped hashes. We assume that a) normal users use benign client programs that are usually authenticated through other mechanisms, and that b) attackers cannot get the plain-text passwords and have to rely on hashes to impersonate a normal user. We can capture the network packets and find out which kind of authentication mechanism is used. The login session that uses those vulnerable authentication mechanisms can then be identified as pass the hash attack. The post-condition is that a malicious process P_5 with user U_1's credential is created. This step has no pre-requisites.

There are three stages during the remote login session. Each stage contains multiple network packets. For example, the second stage, authentication, can be viewed as a sequence made up of client's authentication request, server's challenge, client's challenge response and server's authentication response, as shown in Figure 5. The client first sends a session setup request to the server; then the server responds to the client with a challenge; on receiving the challenge, the client uses the challenge and credentials to do calculations and sends back the result in challenge response packet; finally, the server verifies the result and sends back authentication response indicating whether authentication succeeds or not.



```
No.  Source           Destination      Protocol Length Info
  26 172.29.36.125    172.29.151.231   SMB2        239 Session Setup Request, NTLMSSP_NEGOTIATE
  27 172.29.151.231   172.29.36.125    SMB2        435 Session Setup Response, Error: STATUS_MORE_PROCESSING_REQUIRED, NTLMSSP_CHALLENGE
  28 172.29.36.125    172.29.151.231   SMB2        576 Session Setup Request, NTLMSSP_AUTH, User: CORP\Administrator
  30 172.29.151.231   172.29.36.125    SMB2        151 Session Setup Response
```

Figure 5: A subset of network packets during pass the hash attack.

Pass the Hash Data Generation. The raw data (network packets) are automatically generated by protocol fuzzing. During a pass the hash attack, before a packet is sent to the server, we fuzz certain fields in the application layer (SMB/SMB2 for pass the hash) of the network packets. In this way, (a) the packet structure remains intact, so that the server will not discard the packet; (b) the authentication of pass the hash can be affected; (c) we can get a variety of network packets, and possibly, a variety of network packet sequences so we can get enough diversity in the data for machine learning. The same fuzzing method has also been applied in the generation of benign data from normal network traffic. All the network packets from malicious and benign network traffic are captured using Wireshark. Because of fuzzing, we cannot assure every pass the hash attempt or normal access attempt is successful. For failed pass the hash attempts, we remove them from malicious data. The reason is that, if it fails, the attempt does not generate any impact, so it is not really malicious. For failed normal access, we keep it in benign data, because normal user can also have typos or forget passwords.

Though the raw network logs contain network packets from both malicious and benign network traffic, there are too much redundant data that do not carry useful information for identifying pass the hash, such as timestamp. There are also fields that have fixed values, such as the header length for SMB/SMB2. Values of these two types of fields cannot help identifying pass the hash. Therefore, in the data parser for network logs, such information is removed. Only those that can help identify pass the hash are kept.

Pass the Hash Identification. To identify pass the hash with machine learning, we have two key insights:

- Network communications consist of lots of network packets sent in certain order. What happens at a previous time point can affect what happens afterwards. For example, the first several packets may be a server and a client communicating the protocol to use, and packets afterwards will use the protocol decided;
- Pass the hash relies on certain authentication mechanism to work. Therefore, there must be some differences in certain values in at least one network packet during the network communication.

With these two insights, we decide to build an LSTM neural network that takes a sequence of network packets' types, and outputs the binary label representing whether this sequence is pass the hash network traffic. We have trained a long short-term memory (LSTM) neural network for pass-the-hash identification based on network packets sent between the server (victim) and the client (attacker) during remote logins. Our neural network is presented in Figure 4. M stands for input; H stands for each LSTM block's output; and C stands for each LSTM block's state output. Subscripts stand for the time points, in which h stands for the window size. The network packets' types are defined by the values of fields of interest. If two packets have the same values in all those fields, then the two packets are presented as the same packet type number. Otherwise, different numbers are assigned. In this way, a sequence of network packets can be presented as a sequence of numbers, each standing for a packet of its corresponding packet type.

Now that each network packet is represented as a packet type number, the whole network log can be represented as a whole sequence of packet type numbers. By identifying the start packet for each benign/malicious network communication, we chomp the whole sequence into many variate-length sequences, and the beginning of every sequence is a start packet. Then, we chomp each of those variate-length sequences into one or more fixed-length sequences according to the window size and window shift step size. Depending on whether the sequence comes from a benign traffic or malicious traffic, we assign every fixed-length sequence with the corresponding binary label and get one data sample for our LSTM neural network. If a fixed-length sequence appear in both the benign and malicious data, this sequence is removed from both of them because it cannot help the classification. We then remove duplicate sequences in both the benign data and malicious data. After these two removal processes, all data samples are finally ready to be used.

Data Exfiltration. This technique can be very confidential. It can be encrypted and done through very common protocols like HTTP/HTTPS. As a result, data exfiltration is hard to be identified at the network level. Supposing that attackers are interested in sensitive data, system administrators can enforce disk I/O monitoring with process monitor on sensitive files/folders on a machine. In this way, data exfiltration can be identified at disk level.

6 EVALUATION

At high level, the framework presented can be separated into four phases:

- **Input parsing.** The raw input files are parsed by APT tactic parser, data parsers and system object dependency.
- **APT technique identifying.** APT identifiers use parsed inputs to identify APT techniques.
- **APT tactic matching.** With the help of system object dependencies, the APT tactic matcher matches identified APT techniques into APT tactics.
- **APT tactic ranking.** After tactic matching, the APT tactic ranker ranks all APT tactic instances based on completeness.

We have evaluated each phases and the results show that our framework can correctly detect APT tactics and pick out the one(s) that is fully matched. Specifically, we answer the following questions:

- **RQ1.** How accurately can technique identifiers identify APT techniques?

Zou, Qingtian; Singhal, Anoop; Sun, Xiaoyan; Liu, Peng. "Automatic Recognition of Advanced Persistent Threat Tactics for Enterprise Security." Paper presented at 6th ACM International Workshop on Security and Privacy Analytics 2020, New Orleans, LA, US. March 16, 2020 - March 18, 2020.

Table 3: 384 APT tactics in the repository.

Initial intrusion	Privilege escalation	Credential access	Lateral movement	Impact
Supply chain compromise	Bypass User Account Control	Credential dumping	Pass the hash	Data exfiltration
Exploit public-facing application	DLL search order hijacking	Account manipulation	Logon scripts	Data manipulation
External remote services	New service	Private keys		Data destruction
Spearphishing link	Process injection			Endpoint denial of service

- **RQ2.** Given identified APT techniques, how correctly can the matcher match them into APT tactics and generate instances?
- **RQ3.** Given matched APT tactic instances, can the ranker rank the fully matched APT tactic instances higher than others in the ranked list?
- **RQ4.** How much time and memory does the framework need for each phase?

For evaluation experiments, we first prepared 385 APT tactics in the APT tactic repository. One is a 4-step tactic referred to as tactic001 in Table 4. The other 384 are 5-step tactics presented in Table 3. Each tactic can take any one process belonging to the same technique (column). Therefore, Table 3 provides $4*4*3*2*4 = 384$ APT tactics. Some APT tactics may be very similar at high level. For example, replace one technique in tactic A and get tactic B. The old technique and the new technique are for the same purpose, but done in different ways (e.g. lateral movement by logon scripts or pass the hash). We treat A and B as different APT tactics because different techniques need different technique identifiers to identify them, and those identifiers may use different method and/or system object dependency for identifying.

Out of the 385 APT tactics in the repository, we launched 7 of them in evaluation experiments, presented in Table 4. 3 kinds of logs from those 7 APT tactics are collected, which are network logs (captured by Wireshark), process monitor logs (exported from Prcess Monitor [15]), and windows event logs (exported from Windows Event Viewer). Our test bed, towards which attacks are launched, consists of one Windows AD DC (Windows Server 2012 R2) virtual matching hosting a Windows domain, joined by another Windows 7 virtual machine. The raw input files include 385 APT tactic DOT files, 5.84GB of network logs, 12.6GB of process monitor logs, and 141MB of Windows event logs.

6.1 APT technique identifying

For APT technique identifying, we focus on evaluating one specific technique, which is pass the hash identification. We choose this because other identifiers involve commercial IDS or manually crafted patterns, which have little point in evaluating.

For pass the hash identification, we build an LSTM based neural network. It takes parsed network log files of packets sent between hosts as input, and produce binary results showing whether pass the hash attack is presented in a packet sequence.

With different parameters such as training batch size, LSTM window size and window shift step size, we have trained a total of 144 neural networks. Of all the data fed to the neural network, about 60% are used for training, about 20% are used for validation, and the rest about 20% are used for testing. Every trained neural network is evaluated by false positive rate, false negative rate and F1 score. With different parameters, the number of benign data samples and malicious data samples can also change accordingly. As a result, whether the final dataset is balanced or not is kind of unpredictable, so accuracy is not very helpful. What is more, the true negative samples, which are noises, are not of interest for us, and F1 score does not take true negatives into calculation. Therefore, we choose F1 score as the main criteria.

The best-performing neural network, which has the highest F1 score of 0.9763 on test set, is fed with 2085 benign data samples and 2830 malicious data samples. The false positive rate and false negative rate on the test set of this neural network are 4.437% and 0.252%, respectively. Though the false positive rate is not ideal, we will show that, in tactic matching phase, those false positives cannot produce fully matched APT tactic instance.

Result 1: *The LSTM-based APT technique identifier for pass the hash, which is hard to detect via traditional methods, can identify the technique with low false negative rate, but the false positive rate is not ideal.*

6.2 APT tactic matching

As stated earlier, the number of launched APT tactics in our experiments is small. However, even if there are very few APT tactics happening, the CSOCs need to be aware of all possible APT tactics. With this insight, we evaluated the APT tactic matcher with all the 385 APT tactics in the repository. The goal of evaluating APT tactic matcher is to see whether it can correctly and fully match APT tactic that is actually launched in the following three cases: (a) one attacker is using one APT tactic; (b) multiple attackers are using one same APT tactic; (c) multiple attackers are using multiple different APT tactics. There may be some partially matched APT tactics. This is natural because some APT tactics share the same technique at some point, but the APT tactic matcher just tries its best to match an APT technique into tactic. How much importance should be given to the tactic instance is not evaluated by the APT tactic matcher, but the APT tactic ranker, which will be evaluated in the next subsection.

Case A: One attacker using one APT tactic. In this scenario, tactic001 presented in Table 4 is launched for once in our test bed. The matcher successfully match the target tactic in full. The outputs contains 1 fully matched APT tactic instance of the target tactic and other 123 partially matched instances.

Case B: Multiple attackers using one same APT tactic. In this scenario, tactic001 presented in Table 4 is launched for three times with some different parameters, like PIDs and file names. The

Table 4: A list of APT tactics that were launched.

APT tactic name	1st technique	2nd technique	3rd technique	4th technique	5th technique
tactic001	Supply chain attack	DLL search order hijacking	Logon scripts	Data modification	(None)
tactic002	Supply chain attack	Bypass User Account Control	Credential dumping	Pass the hash	Data exfiltration
tactic003	Supply chain attack	Bypass User Account Control	Credential dumping	Pass the hash	Data modification
tactic004	Supply chain attack	Bypass User Account Control	Credential dumping	Pass the hash	Data destruction
tactic005	Supply chain attack	DLL search order hijacking	Credential dumping	Pass the hash	Data exfiltration
tactic006	Supply chain attack	DLL search order hijacking	Credential dumping	Pass the hash	Data modification
tactic007	Supply chain attack	DLL search order hijacking	Credential dumping	Pass the hash	Data destruction

matcher outputs 9 fully matched APT tactic instances of the target tactic and 375 partially matched instances.

The results show the existence of “duplicates” in fully matched APT tactic instances. The reason behind is that some techniques in the tactic instances can be replaced with others. In Case B, the last two techniques of tactic001 are interchangeable among the 3 attacks. When we launched the 3 attacks, the vulnerable process for DLL search order hijacking remains the same, which made the last two techniques interchangeable. As a result, the final output of fully matched APT tactic instances becomes $3 * 3 = 9$. This is reasonable because at the defender’s side, CSOCs have no idea which attacker is aiming for what in their IT system, so the best practice is to list all possible combinations.

Case C: Multiple attackers using multiple different APT tactics. In this scenario, the APT tactics presented in Table 4 are each launched for once. Note that these APT tactics share some common techniques. The matcher outputs 94 fully matched APT tactic instances and 5468 partially matched instances. The 94 fully matched APT tactic instances include 4 instances of tactic001, 6 of tactic002, 6 of tactic003, 6 of tactic004, 24 of tactic005, 24 of tactic006, and 24 of tactic007.

The similar thing in Case B also happens to Case C. In Case B, the interchangeability results from three attackers using one same tactic; in Case C, the interchangeability results from shared APT techniques among the 7 tactics. For tactic001, its first two techniques (supply chain attack and DLL search order hijacking) are interchangeable among tactic001, tactic005, tactic006, and tactic007, so its fully matched instance number is 4. For tactic002, tactic003 and tactic004, their last two techniques (pass the hash and data exfiltration/modification/destruction) are interchangeable between tactic002 and tactic005, tactic003 and tactic006, and tactic004 and tactic007 respectively. The reason is that pass the hash technique has no prerequisites, so there is no way to chain the first three techniques together with the last two. As a result, the numbers of fully matched instances for tactic002, tactic003 and tactic004 are all $3 * 2 = 6$. For tactic005, tactic006 and tactic007, the first two, the middle one, and the last two techniques are all interchangeable. How the first two and the last two techniques are interchangeable have been discussed earlier; and the middle one technique (credential dumping) is interchangeable among tactic005, tactic006 and tactic007 because they use the same vulnerable process for DLL search order hijacking. Thus, their fully matched instance numbers are all $4 * 3 * 2 = 24$.

One thing worth noting is that after the technique identifying phase, we blindly feed identification results to tactic matching phase. This means that the false positives of pass the hash identification are treated as true positives and used for tactic matching. CSOCs may realize that some of those are false positives, but at the run time, without further inspecting the data, they cannot know that whether the outputs of pass the hash identifier contains false positives or not. In spite of this, the number of fully matched APT tactics is still in consistence with our expectations, which means that, during tactic matching, those false positives cannot be used to produce fully matched APT tactic instances because they do not meet the pre-requisites of system object dependencies.

To further assure the matcher’s resilience to false positives from technique identifying, we carry out an additional experiment. We reproduce Case C, but, this time, we add a filter between the pass the hash identifier and the APT tactic matcher, so that the matcher only gets false positives from pass the hash identifier. Other parts remain the same, and we find out that the matcher only outputs 4 fully matched APT tactic instances of tactic001, and 2126 not fully matched APT tactic instances. Therefore, the APT tactic matcher is resilient to the false positives from technique identifiers.

Result 2: *The APT tactic matcher is resilient to false postives from technique identifiers and can correctly match identified APT techniques into tactics and create APT tactic instances in all three cases.*

6.3 APT tactic ranking

After APT tactic matching, the matcher outputs APT tactic instances, which can be partially of fully matched, to the APT tactic ranker. The APT tactic ranker then ranks these APT tactic instances by completeness. The goal of evaluating APT tactic ranker is to see whether it can correctly pick out fully matched APT tactic instances and put them to the top of the list.

The evaluation results show that, in all of the three Cases A, B and C, the ranker successfully ranks the fully matched instance(s) to the top. In Case B and C, because there are many “duplicate” APT tactic instances with the same 100% completeness, there can be many APT tactic instances, which are the same APT tactics with different system objects, on the top of the ranking list.

Result 3: *The APT tactic ranker can rank the fully matched APT tactic instances on top of the ranked list.*

6.4 Time and memory usage

To evaluate the efficiency of our framework, we have conducted offline measurements of the time and max memory usage for each of the four phases. The machine used for this experiment is a workstation with Intel(R) Xeon(R) E5-2650 v3 processor and 62GB of RAM. The results are presented in Table 5.

Table 5: Max memory and time taken for each phase.

Phase		Max memory taken (MB)	Time taken (s)
Input parsing		38794	1366
APT technique identifying		3812	74
APT tactic matching	Case A	19	1
	Case B	20	1
	Case C	51	13
APT tactic ranking	Case A	16	1
	Case B	17	1
	Case C	49	1

It is shown that, the most resource-consuming phase is the input parsing phase. In this phase, the memory usage can be as high as about twice the size of input files. The memory usage is high because in our Python implementation, we used the data structure of lists a lot. In Python, lists have many redundancies between adjacent elements. To make matters worse, Python does not have any means for memory recycle, which means once some memories are used, they will never be released, until the whole program exits. Therefore, in our Python input parsing implementation, after loading large files into the memories, the program cannot release them after the files' parsing have been finished. If the framework is implemented with other programming languages, like Java, the memory can be better managed.

> ***Result 4:*** *The most resource-consuming phase is input parsing. Other phases consumes acceptable memories and short time. The memory usage and time consumption have the potential to be further reduced.*

7 CONCLUSION AND FUTURE WORK

In this paper, we propose a framework for detecting APT tactics from logs and configuration files. The framework takes previously seen APT tactics, logs and system configuration files as input, and generates a ranked list of APT tactics based on completeness. We also present a detailed case study of a simple 5-step APT tactic, describing how to identify each APT technique in the tactic. Finally, we present the evaluation results of our framework, which clearly shows that the framework can correctly detect APT tactics.

Currently, we implement the framework on one virtual machine. To further improve the efficiency, we plan to implement, validate and evaluate this framework in a cloud environment with multiple virtual machines. Every virtual machine will be dedicated to one task for efficiency. Manual work can be significantly reduced because system administrators only need to care about feeding data into the cloud. The other workloads, including file parsing, technique identifying, and tactic matching will be all completed by the framework automatically.

Our framework matches APT tactics from the repository. However, attackers may update their attack tactics. They may replace old adversary techniques for new ones or add/remove techniques according to their purposes. In the future, we plan to add another component, APT tactic updater, to our framework to automatically handle tactic updating. The updater will take tactic templates from tactic parser, partially matched tactic instances and the pool of unmatched adversary techniques from tactic matcher. Then it can decide when some changes should be made in original tactics, how it should be made, and finally update tactics in the repository.

Another thing worth noticing is that, we simply rank the APT tactics instances based on the completeness. The ranking could be based on more complicated algorithm. For example, CVSS scores can be taken into consideration, and an "impact score" can be calculated for each APT tactic instance. Or another model may be proposed to assess how likely the APT tactic is being used by the attacker. Because this is about APT impact assessment and is out of the scope of our paper, here we do not dig deeper into the ranking problem.

DISCLAIMER

This paper is not subject to copyright in the United States. Commercial products are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

REFERENCES

[1] [n. d.]. Evolving Playbooks in Targeted APT Attacks across Asia Pacific and Japan - Security Boulevard. https://securityboulevard.com/2018/05/evolving-playbooks-in-targeted-apt-attacks-across-asia-pacific-and-japan/

[2] [n. d.]. GitHub - hfiref0x/UACME: Defeating Windows User Account Control. https://github.com/hfiref0x/UACME

[3] [n. d.]. Graphviz - Graph Visualization Software. https://www.graphviz.org/

[4] [n. d.]. Snort - Network Intrusion Detection & Prevention System. https://www.snort.org/

[5] 2019. MITRE ATT&CK™. https://attack.mitre.org [Online; accessed 27. Mar. 2019].

[6] Sean Barnum. 2014. Standardizing cyber threat intelligence information with the Structured Threat Information eXpression (STIX™). *MITRE Corporation* (2014), 1–20.

[7] Saranya Chandran, P. Hrudya, and Prabaharan Poornachandran. 2015. An efficient classification model for detecting advanced persistent threat. In *2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015*. IEEE, 2001–2009. https://doi.org/10.1109/ICACCI.2015.7275911

[8] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. 2017. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1285–1298.

[9] Ibrahim Ghafir, Khaled Rabie, Liangxiu Han, Vaclav Prenosil, Francisco J. Aparicio-Navarro, Robert Hegarty, and Mohammad Hammoudeh. 2018. Detection of advanced persistent threat using machine-learning correlation analysis. *Future Generation Computer Systems* 89 (2018), 349–359. https://doi.org/10.1016/j.future.2018.06.055

[10] Nahid Hossain, Sadegh M Milajerdi, Junao Wang, Birhanu Eshete, Rigel Gjomemo, R Sekar, and Scott Stoller. 2017. SLEUTH : Real-time Attack Scenario Reconstruction from COTS Audit Data. *Proceedings of the 26th USENIX Security Symposium* (2017), 487–504.

[11] Pavlos Lamprakis, Ruggiero Dargenio, David Gugelmann, Vincent Lenders, Markus Happe, and Laurent Vanbever. 2017. Unsupervised detection of APT C&C channels using web request graphs. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10327 LNCS. 366–387. https://doi.org/10.1007/978-3-319-60876-1_17

[12] Kyu Hyung Lee, Xiangyu Zhang, and Dongyan Xu. 2013. High Accuracy Attack Provenance via Binary-based Execution Partition.. In *NDSS*.

[13] Shiqing Ma, Juan Zhai, Fei Wang, Kyu Hyung Lee, Xiangyu Zhang, and Dongyan Xu. 2017. {MPI}: Multiple Perspective Attack Investigation with Semantic Aware Execution Partitioning. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 1111–1128.

[14] Shiqing Ma, Xiangyu Zhang, and Dongyan Xu. 2016. Protracer: Towards Practical Provenance Tracing by Alternating Between Logging and Tainting.. In *NDSS*.

[15] markruss. 2019. Process Monitor - Windows Sysinternals. https://docs.microsoft.com/en-us/sysinternals/downloads/procmon [Online; accessed 26. Aug. 2019].

[16] Sadegh M. Milajerdi, Rigel Gjomemo, Birhanu Eshete, R. Sekar, and V. N. Venkatakrishnan. 2018. HOLMES: Real-time APT Detection through Correlation of Suspicious Information Flows. *2019 IEEE Symposium on Security and Privacy (SP)* (2018). arXiv:1810.01594

[17] Alina Oprea, Zhou Li, Robin Norris, and Kevin Bowers. 2018. MADE: Security Analytics for Enterprise Threat Detection. *ACSAC* (2018). https://doi.org/10.1145/3274694.3274710

[18] Kexin Pei, Zhongshu Gu, Brendan Saltaformaggio, Shiqing Ma, Fei Wang, Zhiwei Zhang, Luo Si, Xiangyu Zhang, and Dongyan Xu. 2016. Hercule: Attack story reconstruction via community discovery on correlated log graph. In *Proceedings of the 32Nd Annual Conference on Computer Security Applications*. ACM, 583–595.

[19] Joseph Sexton, Curtis Storlie, and Joshua Neil. 2015. Attack chain detection. *Statistical Analysis and Data Mining* (2015). https://doi.org/10.1002/sam.11296

[20] Sana Siddiqui, Muhammad Salman Khan, Ken Ferens, and Witold Kinsner. 2016. Detecting Advanced Persistent Threats using Fractal Dimension based Machine Learning Classification. In *Proceedings of the 2016 ACM on International Workshop on Security And Privacy Analytics - IWSPA '16*. ACM Press, New York, New York, USA, 64–69. https://doi.org/10.1145/2875475.2875484

[21] G Zhao, K. Xu, L Xu, and B. Wu. 2015. Detecting APT malware infections based on malicious DNS and traffic analysis. *IEEE Access* 3 (2015), 1132–1142. https://doi.org/10.1109/ACCESS.2015.2458581

The Impossibility of Efficient Quantum Weak Coin-Flipping

Carl A. Miller

Joint Center for Quantum Information and Computer Science
University of Maryland, College Park, MD 20742, USA

National Institute of Standards and Technology,
100 Bureau Dr., Gaithersburg, MD 20899, USA

Abstract

How can two parties with competing interests carry out a fair coin flip, using only a noiseless quantum channel? This problem (quantum weak coin-flipping) was formalized more than 15 years ago, and, despite some phenomenal theoretical progress, practical quantum coin-flipping protocols with vanishing bias have proved hard to find. In the current work we show that there is a reason that practical weak quantum coin-flipping is difficult: any quantum weak coin-flipping protocol with bias ϵ must use at least $\exp(\Omega(1/\sqrt{\epsilon}))$ rounds of communication. This is a large improvement over the previous best known lower bound of $\Omega(\log\log(1/\epsilon))$ due to Ambainis from 2004. Our proof is based on a theoretical construction (the two-variable profile function) which may find further applications.

1 Introduction

Suppose that Alice and Bob are two cooperating but mutually mistrustful parties, and they must make a unified decision between two choices (X and Y). Alice wants choice X, and Bob wants choice Y. However, neither of them will gain if they do not agree on their decision. How can the decision be made fairly? A natural solution would be for Alice and Bob to have a trusted third party (Charlie) flip a coin and report the result to both Alice and Bob. But, can coin-flipping be done in absence of any trusted third party or common source of randomness?

In this paper we will be concerned with the question of whether coin-flipping can be done if Alice and Bob share a two-way noiseless quantum channel. A standard way to model a protocol in this scenario is like so (see Figure 1). Let n be an even positive integer.

1. Alice possesses a quantum system $\mathcal{A}$, which she controls, and Bob possesses a quantum system $\mathcal{B}$ which he controls.
2. There is an additional quantum system $\mathcal{M}$, initially possessed by Alice, which stores quantum messages exchanged by Alice and Bob during the protocol.
3. If i is odd, then on the ith round of the protocol, Alice performs a prescribed joint quantum operation on $\mathcal{A}$ and $\mathcal{M}$, and then sends $\mathcal{M}$ across the quantum channel to Bob.
4. If i is even, then on the ith round of the protocol, Bob performs a prescribed joint quantum operation on $\mathcal{B}$ and $\mathcal{M}$ and then sends $\mathcal{M}$ across the channel to Alice.
5. After the nth round of communication, Alice performs a binary measurement on $\mathcal{A}$ and reports the result as a bit (a), and Bob performs a binary measurement on $\mathcal{B}$ and reports the result as a bit (b).

We suppose that Alice desires the outcome $a = b = 0$, and that Bob desires the outcome $a = b = 1$. It is presumed that an "honest" party will carry out their operations and measurements exactly as prescribed; however, a dishonest party may perform arbitrary manipulations of the quantum systems that they possess, and may perform any final measurement that they choose at the end of the protocol. (In particular, there is no restriction on the computational power of Alice or Bob.)

We say such a protocol is a **weak coin-flipping protocol with bias** ϵ if the following hold:

Miller, Carl A. "The Impossibility of Efficient Quantum Weak Coin-Flipping." Paper presented at 52nd Annual ACM Symposium on Theory of Computing (STOC 2020), Chicago, IL, US. June 22, 2020 - June 26, 2020.

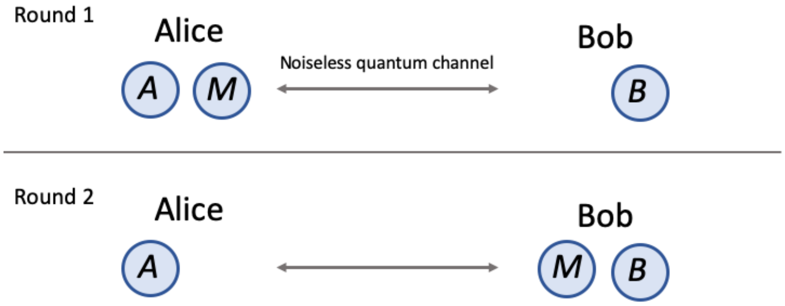


Figure 1: The first two rounds of a weak quantum coin-flipping protocol.

1. If Alice and Bob both perform honestly, then $\mathbb{P}(a = b = 0)$ is exactly $\frac{1}{2}$ and $\mathbb{P}(a = b = 1)$ is exactly $\frac{1}{2}$.
2. If Alice behaves dishonestly and Bob behaves honestly, then $\mathbb{P}(b = 0) \leq \frac{1}{2} + \epsilon$.
3. If Bob behaves dishonestly and Alice behaves honestly, then $\mathbb{P}(a = 1) \leq \frac{1}{2} + \epsilon$.

These conditions assert that Alice cannot bias the outcome by more than ϵ in her favor, and Bob cannot bias the result by more than ϵ in his favor. (A **strong coin-flipping protocol with bias** ϵ is one which guarantees that a dishonest party cannot bias the result of the coin flip by more than ϵ in either direction. Strong coin-flipping with vanishing bias is impossible by an elementary argument attributed to A. Kitaev — see [12].)

We note that in a classical analogue of the setting described above — i.e., where $\mathcal{A}, \mathcal{B}, \mathcal{M}$ are random variables, Alice's and Bob's operations are stochastic maps, and the two parties are computationally unlimited — it is elementary to show that any protocol allows either Alice to force the outcome to always be 0, or Bob to force the outcome to always be 1. (Thus, achieving bias less than $\frac{1}{2}$ is impossible.) This fact can be overcome by putting computational restrictions on Alice and Bob, and classical coin-flipping is itself an elegant and extensively studied topic (see [6, 20]). The main motivation to study quantum coin-flipping is that, in contrast to classical coin-flipping, it allows security proofs based on physical assumptions only.

Quantum coin-flipping was formalized as early as 1998 ([15]), and a series of works proved coin-flipping protocols with progressive improvements in the bias. Aharonov et al. [2] proved bias 0.42. Spekkens and Rudolph, and independently Ambainis, proved successive results [22, 23, 4] which brought the bias down to $\frac{\sqrt{2}-1}{2} \approx 0.207$. These results involved a small constant number of rounds of communication. Mochon [17, 18] then introduced a family of quantum weak coin-flipping protocols which approach bias $1/6 \approx 0.166$, with the number of communication rounds tending to infinity.

Finally, in a landmark work in 2007, Mochon [19] showed the existence of a family of weak coin-flipping protocols with bias tending to zero. Mochon exploited the idea of *point games* (a concept also attributed to A. Kitaev) to achieve this result. Mochon's existence proof was later simplified, re-written and published by Aharonov et al. [1]. Then, in recent work [5], Arora et al. introduced an algorithm which effectively constructs the protocols in the family whose existence was proven by Mochon.

Following this phenomenal progress, at least one major loose end remains. The number of communication rounds used in the protocols in [19, 1] was only shown to be $(1/\epsilon)^{O(1/\epsilon)}$. This asymptotic quantity is hardly efficient or practical. Meanwhile, the best known *lower* bound on the number of communication rounds [4] is $\Omega(\log\log(1/\epsilon))$, leaving a vast range of uncertainty about the optimal resources needed to achieve vanishing bias.

How many rounds of quantum communication are needed to achieve a particular bias ϵ? For example, in a different setting (strong classical coin-flipping with computational hardness assumptions), Cleve [11] showed that the number of communication rounds was $\Omega(1/\epsilon)$, and this bound was shown to be achievable [20]. Could a similar relationship exist for quantum weak coin-flipping?

1.1 Summary of result

In the current paper, we prove the following lower bound on the number of communication rounds for quantum weak coin-flipping (see Theorem 8.2):

Theorem 1.1. *Let* $\mathbf{C}$ *be an* n*-round quantum weak coin-flipping protocol with bias* ϵ*. Then,*

$$n \quad \geq \quad \exp\left(\Omega\left(\frac{1}{\sqrt{\epsilon}}\right)\right). \tag{1}$$

This result shows that, at least in the standard model, practical quantum weak coin-flipping with vanishing bias is not feasible.

The proof of this result builds on previous techniques, including the concept of a *valid time-independent point game* (abbreviated as "valid TIPG"). A valid TIPG is a pair of real-valued functions m_1, m_2 on $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ that satisfy a certain infinite set of linear constraints (see subsections 3.2–3.3). It is known that any weak coin-flipping protocol determines a valid TIPG, and vice versa. This correspondence was used to prove the family of protocols with vanishing bias in [19]. Here I prove a negative result: any TIPG obtained from a weak coin-flipping protocol with small bias must have very large 1-norm — i.e., $\|m_1\|_1 + \|m_2\|_1$ must be very large as a function of ϵ. Since there is a relationship between the number of communication rounds of a protocol and the 1-norms of its associated time-independent point games, this implies the main result.

Ambainis's original bound [4] of $\Omega(\log\log(1/\epsilon))$ was based on a more direct study of quantum weak coin-flipping protocols: he performed an inductive argument, using the fidelity function, on the intermediate states arising in the protocol. This approach appears to be fairly different from the one in this paper, although it may be possible to relate the two.

Theorem 1.1 is a further step in mapping out the full range of cryptographic possibilities in a two-party quantum setting (see [7, 24] for surveys on this topic). A number of other negative results are known: secure two-party computation (under certain definitions) is impossible [13, 8], and strong coin-flipping [12, 9], bit commitment [14, 16, 9], and oblivious transfer [13, 10] are all impossible except with fixed positive bias. This paper shows that the case of quantum weak coin-flipping is different: it can be achieved with arbitrarily small bias, but it is impossible to do so in polynomial time.

Certainly, this is not the end of the story. The model for quantum weak coin flipping makes a number of assumptions, including that the players exchange information in discrete stages, and that they are completely unconstrained in their ability to manipulate any quantum systems that are not under the control of the other player. This impossibility result gives us additional motivation to study coin-flipping in other settings, including relativistic models.

1.2 Outline

Sections 2–3 of this paper cover preliminaries and known material, and then new contributions appear in sections 4–8. In section 4, I define the *profile* of a time-independent point game, which is a two-variable function associated to a time-independent point game that distills some of its most relevant information. Sections 5–6 prove some mathematical lemmas, including a key result on the behavior of highly concentrated rational functions (Proposition 5.3). Section 7 proves the result about the 1-norm of a time-independent point game using tools from the previous sections. Finally, section 8 proves Theorem 1.1. I conclude by noting some further directions in section 9.

1.3 Acknowledgements

I am grateful to Aarthi Sundaram (my co-author on a companion project) for many interesting discussions about the coin-flipping literature which helped to seed some of the ideas contained here. This paper also owes a large debt to Alexandre Eremenko, who showed me the complex analysis method that I used to prove Proposition 5.3. Thanks to Gorjan Alagic, Jonathan Katz, Yi-Kai Liu, and Scott Wolpert for their help with this project, and to Michael Newman for giving me an introduction to [1] some years ago.

This work is a contribution of the U. S. National Institute of Standards and Technology (NIST), and is not subject to copyright in the United States. Any mention of commercial products is for information purposes only, and does not imply endorsement by NIST.

2 Preliminaries

Let $\mathbb{R}$ denote the set of real numbers, and let $\mathbb{R}_{\geq 0}$ denote the set of nonnegative real numbers. If a, b are real numbers with $a \leq b$, then $[a, b]$ denotes the closed interval $\{x \mid a \leq x \leq b\}$, (a, b) denotes the open interval $\{x \mid a < x < b\}$, and $[a, b)$ and $(a, b]$ are similarly defined. We let ∞ denote infinity, and define intervals such as $[a, \infty] \subseteq \mathbb{R} \cup \{\infty\}$ in the obvious way. If A is a set and $B \subseteq A$ is a subset, then $A \smallsetminus B$ denotes the set of all elements of A that are not in B.

Our notation follows previous work [19, 1, 5] in part. Since we will work extensively with functions on $\mathbb{R}_{\geq 0}$ that have finite support, we make the following definitions.

Definition 2.1. *For any $x \in \mathbb{R}_{\geq 0}$, let $[\![x]\!]$ denote the function from $\mathbb{R}_{\geq 0}$ to $\mathbb{R}$ which maps x to 1 and is zero elsewhere. For any $x, y \in \mathbb{R}_{\geq 0}$, let $[\![x, y]\!]$ denote the function from $\mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0}$ to $\mathbb{R}$ which maps (x, y) to 1 and is zero elsewhere.*

If f is a real-valued function on a set S, define $f^+ \colon S \to \mathbb{R}$ and $f^- \colon S \to \mathbb{R}$ by

$$f^+(x) = \max\{0, f(x)\}, \tag{2}$$
$$f^-(x) = \max\{0, -f(x)\}. \tag{3}$$

Note that $f = f_+ - f_-$. Let Supp f denote the support of f (i.e., the set of points in S on which f is nonzero). Let $f^\top$ denote the function $f^\top(x, y) = f(y, x)$. When f is a function with finite support, then (even if the set S is not countable) we will write $\sum_{s \in S} f(s)$ to mean the sum of $f(s)$ over all points in the support of f. The expression $\|f\|_1$ denotes the sum $\sum_{s \in S} |f(s)|$ (that is, the 1-norm of f).

The function $\log \colon \mathbb{R}_{\geq 0} \to \mathbb{R} \cup \{-\infty\}$ denotes the logarithm in base 2.

We use the term **universal function** to mean a function that is not dependent on any variables other than its input variables. Thus, even if we refer to a universal function after some variables have been quantified (e.g., "for all c, ...") it is understood that the function has no implicit dependencies on those variables. We will use boldface Roman letters $(\mathbf{A}, \mathbf{B}, \ldots)$ for universal functions.

When we use asymptotic big-O notation, we may use $O(u)$ as a set (e.g., "there exists $\mathbf{F}(u) \in O(u)$ such that ...") or as a placeholder for a function (e.g., "$x = y + O(z)$"). When a big-O expression is used as a placeholder, it is understood that it also represents a universal function with no implicit dependencies.

If Q is an event, then we write $\mathbb{P}[Q]$ for the probability of Q, and if X is a real-valued random variable, then we write $\mathbb{E}[X]$ for the expectation of X. A **stochastic map** from a set A to a set B is an indexed set of nonnegative real values

$$V = \{v_{ab} \mid a \in A, b \in B\} \tag{4}$$

such that $\sum_b v_{ab} = 1$ for all $a \in A$. For any $a \in A$, the values $\{v_{ab} \mid b \in B\}$ define a random variable on B which we denote by $V(a)$.

If $\mathcal{A}$ and $\mathcal{B}$ are Hilbert spaces, then we may write $\mathcal{AB}$ for the tensor product $\mathcal{A} \otimes \mathcal{B}$.

2.1 Complex analysis

We briefly cover some complex analysis tools that will be important in section 5. The reader can consult [3] for more details.

Let $\mathbb{C}$ denote the set of complex numbers. We will apply addition and multiplication to $\mathbb{C} \cup \{\infty\}$ using natural rules ($c + \infty = \infty, 1/\infty = 0$, etc.). For any $z \in \mathbb{C}$ and $r \geq 0$, let

$$\mathbb{D}(z, r) = \{w \in \mathbb{C} \mid |z - w| < r\} \tag{5}$$
$$\mathbb{S}(z, r) = \{w \in \mathbb{C} \mid |z - w| = r\}. \tag{6}$$

(The set $\mathbb{D}(z, r)$ is the open disc of radius r centered at z, and the set $\mathbb{S}(z, r)$ is the circle of radius r centered at z.) When $z = 0$ and $r = 1$, we may write these sets simply as $\mathbb{D}$ and $\mathbb{S}$. Also let

$$\mathbb{H} = \{w \in \mathbb{C} \mid \text{Im } w > 0\}. \tag{7}$$

If $Y \subseteq \mathbb{C} \cup \{\infty\}$, then $\overline{Y}$ denotes the closure of Y. For ease of notation we will write $\overline{\mathbb{D}}(z, r)$ for the closure of $\mathbb{D}(z, r)$. Let $\overline{\mathbb{H}}$ denote the set $\{z \in \mathbb{C} \mid \text{Im } z \geq 0\} \cup \{\infty\}$.

If $S \subseteq \mathbb{C}$ is an open set, then a function $f \colon S \to \mathbb{C}$ is **analytic** if for any $p \in S$, f can be expressed as a power series on some open neighborhood of p. If f is analytic and $\overline{\mathbb{D}}(z, r)$ is a closed disc within its domain, then the following equation always holds:

$$\frac{1}{2\pi} \int_0^{2\pi} f(z + re^{i\theta}) d\theta \quad = \quad f(z). \tag{8}$$

3 Review of quantum weak coin-flipping

In this section we review the common formal framework for quantum weak coin-flipping, including the mathematical construction of a *point game* (which is attributed to A. Kitaev). Since this framework has already seen thorough treatment in [19, 1, 5], we will mainly provide only definitions and statements of results here. Our terminology and notation are derived most directly from [1].

3.1 Weak coin flipping protocols

The definition of a weak coin flipping protocol (which we will first sketch, and then state formally) is intended to capture a general situation where two parties with competing interests are trying to fairly flip a coin. There are two possible outcomes: 0 (or "heads") which is the desired outcome for Alice, and 1 (or "tails") which is the desired outcome for Bob. There is no trusted third party in the protocol, and thus it consists entirely of communication between Alice and Bob. At the end of the protocol, the parties report bits a and b respectively (representing what they ostensibly believe to be the outcome of the coin flip).

The protocol is accomplished by Alice and Bob passing a quantum system (represented by the finite-dimensional Hilbert space $\mathcal{M}$) back and forth between them, while keeping private systems (represented by $\mathcal{A}$ and $\mathcal{B}$) to themselves. In each odd round i, Alice performs a unitary operator U_i on $\mathcal{AM}$ followed by a binary projective measurement $\{E_i, I_{\mathcal{AM}} - E_i\}$ on $\mathcal{AM}$. If the latter measurement fails — that is, if its postmeasurement state is not in Supp E_i — then Alice aborts the protocol and simply reports her favored outcome 0. (This event is understood to mean that Alice has stopped because she suspects cheating.) On even rounds, Bob does analogous operations with operators U_i, E_i on $\mathcal{MB}$. At the end of the protocol, if neither party has aborted, they each perform a binary projective measurement on their private system and report the result (as a and b, respectively). The definition requires that if Alice and Bob behave honestly in the protocol, then the probability that $a = b = 0$ is 1/2, and the probability that $a = b = 1$ is 1/2.

Definition 3.1. *A **weak coin-flipping protocol** $\mathbf{C}$ consists of the following data:*

- *Finite-dimensional Hilbert spaces $\mathcal{A}, \mathcal{M}, \mathcal{B}$ (Alice's system, the message system, and Bob's system),*
- *An even positive integer n (the number of rounds),*
- *An initial pure state ψ_0 on $\mathcal{AMB}$ of the form*

$$\psi_0 \quad = \quad \psi_{\mathcal{A},0} \otimes \psi_{\mathcal{M},0} \otimes \psi_{\mathcal{B},0}, \tag{9}$$

- *For each odd value i from $\{1, 2, \ldots, n\}$, a unitary operator U_i on $\mathcal{AM}$ and a Hermitian projection operator E_i on $\mathcal{AM}$,*
- *For each even value i from $\{1, 2, \ldots, n\}$, a unitary operator U_i on $\mathcal{MB}$ and a Hermitian projection operator E_i on $\mathcal{MB}$,*
- *Binary projective measurements $\{\Pi^0_{\mathcal{A}}, \Pi^1_{\mathcal{A}}\}$ and $\{\Pi^0_{\mathcal{B}}, \Pi^1_{\mathcal{B}}\}$ on $\mathcal{A}$ and $\mathcal{B}$, respectively,*

where the following condition holds: the state

$$\psi_n \quad := \quad E_n U_n E_{n-1} U_{n-1} \cdots E_1 U_1 \psi_0 \tag{10}$$

*(which is referred to as the **final state** of the protocol) satisfies*

$$\left\| \Pi^0_{\mathcal{A}} \otimes \Pi^0_{\mathcal{B}} \left| \psi_n \right\rangle \right\|^2 = \left\| \Pi^1_{\mathcal{A}} \otimes \Pi^1_{\mathcal{B}} \left| \psi_n \right\rangle \right\|^2 = \tfrac{1}{2}. \tag{11}$$

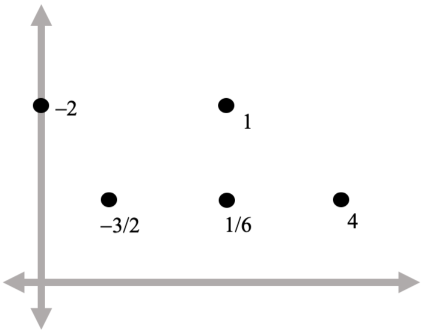


Figure 2: An example of a diagram of a 2-dimensional move w on a Cartesian coordinate system. The marked points are the elements of Supp w, and the labels indicate the values taken by w.

For the definition above, the states

$$\psi_i \quad := \quad E_i U_i E_{i-1} U_{i-1} \cdots E_1 U_1 \psi_0 \tag{12}$$

for $i \in \{1, \ldots, n-1\}$, are referred to as the **intermediate states** of the protocol.

Let us suppose that Bob (whose goal is to force Alice to report $a = 1$) chooses to behave dishonestly in the protocol. In that case, he can apply arbitrary unitary operations $V_2, V_4, V_6, \ldots$ on $\mathcal{MB}$ in place of $E_2U_2, E_4U_4, E_6U_6, \ldots$. (We do not account for any measurements performed by Bob in this case, since his own output is irrelevant.) This motivates the following definition. The **cheating probability** for Bob (in protocol $\mathbf{C}$) is the maximum of

$$\left\| \Pi^1_{\mathcal{A}} V_n (E_{n-1}U_{n-1}) V_{n-2} (E_{n-3}U_{n-3}) V_{n-4} (E_{n-5}U_{n-5}) \ldots V_2 (E_1 U_1) \left| \psi_0 \right\rangle \right\|^2 \tag{13}$$

over all unitary operators $V_2, V_4, \ldots, V_n$ on $\mathcal{MB}$.

Let P_B^* denote the cheating probability for Bob, and let P_A^* denote the cheating probability for Alice (defined analogously). Then, the **bias** of the weak coin-flipping protocol $\mathbf{C}$ is the quantity

$$\max \left\{ P_A^* - \frac{1}{2}, P_B^* - \frac{1}{2} \right\}. \tag{14}$$

3.2 Valid point games

Valid point games are elegant mathematical constructions which, as we will see, are in a near-perfect correspondence with weak coin-flipping protocols. Because they are a lot simpler to define, valid point games provide a convenient method of reduction for questions about weak coin-flipping protocols. We will give the definition of valid point games in this subection, and then explain their relationship to weak coin-flipping protocols in subsection 3.3. We use standard terminology with a few additions.

Definition 3.2. *A **one-dimensional move** is a function $\ell \colon \mathbb{R}_{\geq 0} \to \mathbb{R}$ that has finite support. A one-dimensional move is called a one-dimensional **configuration** if $\ell \geq 0$. A **two-dimensional move** is a function $q \colon \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ that has finite support. A two-dimensional move is called a two-dimensional **configuration** if $q \geq 0$.*

Remark 3.3. *When we use the words **move** or **configuration** by themselves, we will always mean a two-dimensional move or two-dimensional configuration.*

Given a move q, it can be helpful to visualize q by graphing its support set Supp q and writing out the values $q(x, y)$ associated to each point $(x, y) \in$ Supp q. See Figure 2 for an example of a move whose support is of size 5.

A time-dependent point game is, roughly speaking, a finite sequence of two-dimensional configurations $z_0, \ldots, z_n$. However, for mathematical convenience, we define time-dependent point games in terms of the moves $(z_i - z_{i-1})$ rather than the configurations z_i. Recall that for any real-valued function f, we write f^+ and f^- for the positive and negative parts of f, respectively.

Definition 3.4. *A time-dependent point game M is a sequence of moves $(m_1, \ldots, m_n)$ such that*

$$\left(\sum_{i=1}^{n} m_i \right)^- + \sum_{i=1}^{j} m_i \quad \geq \quad 0 \tag{15}$$

for all $j \in \{1, 2, \ldots, n-1\}$.

In the above definition, we refer to $\left(\sum_{i=1}^n m_i\right)^-$ as the **initial configuration** of M, and we refer to $\left(\sum_{i=1}^n m_i\right)^+$ as the **final configuration** of M. The configurations on the left side of inequality (15) are referred to as the **intermediate configurations** of M. The **support** of a point game M, denoted Supp M, is the union of the supports of the moves in M.

Next we will define valid moves.

Definition 3.5. *A one-dimensional move ℓ is **valid** if the following conditions hold:*

$$\sum_x \ell(x) \quad = \quad 0 \tag{16}$$

$$\sum_x \left(\frac{x}{x+\lambda} \right) \ell(x) \quad \geq \quad 0 \qquad \forall \lambda > 0 \tag{17}$$

$$\sum_x x \cdot \ell(x) \quad \geq \quad 0. \tag{18}$$

(We note that condition (18) is actually redundant, since it can be proved from (17).) Equivalently, a move is valid if and only if its inner product with any operator monotone function is nonnegative (see subsection 3.2 of [1]). We note the following useful fact, which is easily proved from Definition 3.5.

Proposition 3.6. *If ℓ is a valid one-dimensional move, and $c > 0$, then the function $x \mapsto \ell(cx)$ is also a valid one-dimensional move.* $\square$

For any two-dimensional move q, the **rows** of q are the functions of the form $x \mapsto q(x, y)$ (for $y \in \mathbb{R}_{\geq 0}$) and the **columns** of q are the functions of the form $y \mapsto q(x, y)$ (for $x \in \mathbb{R}_{\geq 0}$).

Definition 3.7. *A two-dimensional move is **horizontally valid** if its rows are valid. A two-dimensional move is **vertically valid** if its columns are valid. A point game $(m_1, \ldots, m_n)$ is a **valid point game** if m_i is horizontally valid for all odd i, and m_i is vertically valid for all even i.*

For later use, we define the concept of a time-independent point game, which is simply a sequence of two-dimensional moves with no restrictions on nonnegativity (unlike Definition 3.4). In this paper, as in previous papers on weak coin-flipping, it is only useful to consider time-independent point games that involve 2 moves, and so we confine our definition accordingly.

Definition 3.8. *A **time-independent point game (TIPG)** is a pair of moves (r_1, r_2).*

We apply the term "valid" to time-independent point games in the obvious way: a time-independent point game (m_1, m_2) is valid if m_1 is horizontally valid and m_2 is vertically valid. Note that any time-dependent point game $(m_1, \ldots, m_n)$ yields a time-independent point game

$$(m_1 + m_3 + m_5 + \cdots, m_2 + m_4 + m_6 + \cdots) \tag{19}$$

and the latter game is valid if $(m_1, \ldots, m_n)$ is valid.

Remark 3.9. *When we use the term **point game** by itself, we will always mean a time-dependent point game.*

3.3 The relationship between point games and coin-flipping protocols

Now we will state the known results which motivate our study of valid point games. There is a close correspondence between valid point games and weak coin-flipping protocols, and this correspondence allows us to deduce assertions about coin-flipping protocols (of both existence and impossibility) by studying properties of point games.

Theorem 3.10. *Suppose that* $\mathbf{C}$ *is an* n*-round weak coin-flipping protocol with cheating probabilities* P_A^* *and* P_B^**, and that* $\delta > 0$*. Then, there exists a valid point game* $M = (m_1, \ldots, m_n)$ *with initial configuration* $\frac{1}{2}([\![1,0]\!] + [\![0,1]\!])$ *and final configuration* $[\![P_A^* + \delta, P_B^* + \delta]\!]$*.*

If M is a point game from the configuration $\frac{1}{2}([\![0,1]\!] + [\![1,0]\!])$ to a single point $[\alpha, \beta]$, then we will naturally refer to the quantity $\max\{\alpha, \beta\} - 1/2$ as the **bias** of M. The above theorem can be understood as asserting that if an n-round weak coin-flipping protocol exists with bias ϵ, then there are n-round valid point games with bias arbitrarily close to ϵ. A proof of Theorem 3.10, which is based on semidefinite programming duality, is given in [1].[1] (We note that Theorem 3.10 also has a converse, although we will not need it here. See Theorem 3 and Theorem 4 in [1].)

Next we assert a theorem about the relationship between weak coin-flipping protocols and time-independent point games.

Theorem 3.11. *Suppose that* $\mathbf{C}$ *is an* n*-round weak coin-flipping protocol with cheating probabilities* $\alpha := P_A^*$ *and* $\beta := P_B^*$*, and suppose that* $\delta > 0$*. Then, there exists a valid TIPG* $R = (r_1, r_2)$ *such that*

$$r_1 + r_2 \quad = \quad -\frac{1}{2}[\![0,1]\!] - \frac{1}{2}[\![1,0]\!] + [\![\alpha + \delta, \beta + \delta]\!] \tag{20}$$

and

$$\|r_1\|_1 + \|r_2\|_1 \quad \leq \quad 2n. \tag{21}$$

Proof. By Theorem 3.10, there is a valid time-dependent point game $M = (m_1, \ldots, m_n)$ whose sum is equal to the right-hand side of equation (20). Note that since the initial configuration $\frac{1}{2}[\![0,1]\!] + \frac{1}{2}[\![1,0]\!]$ of this game has 1-norm equal to one, and each move m_i sums to zero, the intermediate configurations all also have 1-norm equal to one, and therefore $\|m_i\|_1 \leq 2$ for all i. Therefore the TIPG $(r_1, r_2) := (m_1 + m_3 + m_5 + \cdots, m_2 + m_4 + m_6 + \cdots)$ satisfies the desired conditions. □

We make one final note about symmetry in valid TIPGs.

Remark 3.12. *Let us say that a move* q *is* ***symmetric*** *if* $q^\top = q$*, and that a TIPG* $Q = (q_1, q_2)$ *is symmetric if* $q_2 = q_1^\top$*. Note that if a valid TIPG* $R = (r_1, r_2)$ *achieves a symmetric move (such as the biased coin-flip move* $[\frac{1}{2} + \epsilon, \frac{1}{2} + \epsilon] - \frac{1}{2}[0,1] - \frac{1}{2}[1,0]$*) then there necessarily exists a symmetric valid TIPG that achieves the same move — specifically,* $R' := ((r_1 + r_2^\top)/2, (r_1^\top + r_2)/2)$*.*

4 The profile of a move

We now begin the contributions of this paper. We start by introducing the idea of a **profile function** of a move. If q is a two-dimensional move, then its profile function, denoted $\widehat{q}$, is a real-valued function on $[1, \infty] \times [1, \infty]$. Profile functions have geometric features that we can use to our advantage when trying to answer questions about point games.

4.1 Definition

We begin with the definition of the profile function in the one-dimensional case. The one-dimensional profile function can be thought of simply as a construction that bundles together the three conditions that define a valid move (Definition 3.5).

Definition 4.1. *For any positive real number* x*, let* $P_x \colon [1, \infty] \to \mathbb{R}$ *be defined by*

$$P_x(\alpha) \quad = \quad \begin{cases} 1 & \text{if } \alpha \in [1,2) \\ (x\alpha - x)/(x + \alpha - 2) & \text{if } \alpha \in [2, \infty) \\ x & \text{if } \alpha = \infty. \end{cases} \tag{22}$$

[1] The authors in [1] define the class of "EBM point games" and effectively prove (in their Proposition 2.5) that our Theorem 3.10 holds with the phrase "valid point game" replaced by "EBM point game." Then, in section 3, they prove that every EBM point game is a valid point game.

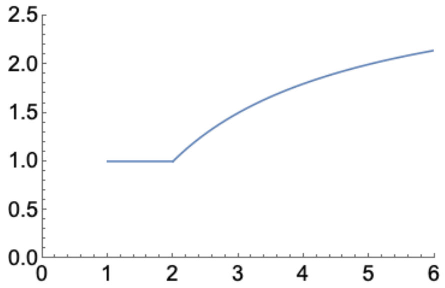


Figure 3: A graph of the function $P_3(\alpha)$. (Mathematica)

If $x = 0$, then let $P_0 \colon [1, \infty] \to \mathbb{R}$ be defined by

$$P_0(\alpha) = \begin{cases} 1 & \text{if } \alpha \in [1, 2) \\ 0 & \text{if } \alpha \in [2, \infty]. \end{cases} \tag{23}$$

*The function P_x is referred to as the **profile of** x.*

*For any one-dimensional move $\ell \colon \mathbb{R}_{\geq 0} \to \mathbb{R}$, the **profile of** ℓ, denoted $\widehat{\ell}$, is the function from $[1, \infty]$ to $\mathbb{R}$ given by*

$$\widehat{\ell} = \sum_x \ell(x) P_x. \tag{24}$$

The profile function $\widehat{\ell}$ collects some useful information about ℓ. One can easily verify from the definition that

$$\begin{aligned} \widehat{\ell}(1) &= \sum_x \ell(x) & (25) \\ \widehat{\ell}(2) &= \sum_{x>0} \ell(x) & (26) \\ \widehat{\ell}(\infty) &= \sum_x x \cdot \ell(x). & (27) \end{aligned}$$

For illustration, we give a graph of the function $P_3(\alpha)$ (that is, the profile of $[\![3]\!]$) in Figure 3.

Proposition 4.2. *A one-dimensional move ℓ is valid if and only if $\widehat{\ell}$ is nonnegative and $\widehat{\ell}(1) = 0$.*

Proof. Given Definition 3.5, this follows from equations (25) and (27) together with the observation that

$$\sum_x \left(\frac{x}{\lambda + x} \right) \ell(x) \geq 0 \qquad \Longleftrightarrow \qquad \widehat{\ell}\,(\lambda + 2) \geq 0. \tag{28}$$

This completes the proof. □

The reader will observe that there are multiple ways that the one-dimensional profile function could have been defined to achieve the property in Proposition 4.2 (e.g., using different ranges for α and different rational expressions). This particular choice of definition will make some of the arguments in section 7 mathematically easier. One of the reasons that this definition is convenient is that the profile of $[\![1]\!]$ is simply the constant function $\alpha \mapsto 1$.

Next we define the profile of a two-dimensional move.

Definition 4.3. *For any two-dimensional move q, the **profile of** q, denoted $\widehat{q}$, is the function from $[1, \infty] \times [1, \infty]$ to $\mathbb{R}$ given by*

$$\widehat{q}(\alpha, \beta) = \sum_{x,y} q(x, y) P_x(\alpha) P_y(\beta). \tag{29}$$

As with the one-dimensional profile, some of the values of a two-dimensional profile $\widehat{q}$ have natural expressions — for example,

$$\widehat{q}(1, \infty) = \sum_{x,y} y \cdot q(x, y) \tag{30}$$

and

$$\widehat{q}(\infty, \infty) = \sum_{x,y} x \cdot y \cdot q(x, y). \tag{31}$$

The basic motivation to study the two-dimensional profile function is the following proposition.

Proposition 4.4. *If q is a move that is either horizontally valid or vertically valid, then $\widehat{q} \geq 0$.*

Proof. Suppose that q is horizontally valid. Then, the single-variable profiles of the rows of q are all nonnegative. Thus for any $\alpha, \beta \in [1, \infty]$,

$$\begin{aligned} \widehat{q}(\alpha, \beta) &= \sum_y \left(\sum_x q(x, y) P_x(\alpha) \right) P_y(\beta) && (32) \\ &\geq 0. && (33) \end{aligned}$$

The vertically valid case is similar. □

As a consequence of the above proposition, if $M = (m_1, \ldots, m_n)$ is a valid time-dependent point game and $z_0, z_1, \ldots, z_{n-1}, z_n$ are its configurations, we must have

$$\widehat{z_0}(\alpha, \beta) \leq \widehat{z_1}(\alpha, \beta) \leq \ldots \leq \widehat{z_n}(\alpha, \beta) \tag{34}$$

for any $\alpha, \beta \in [1, \infty]$. The profile function gives us an infinite family of constraints that must be satisfied by the initial and final configurations of any valid point game.

We make note of some additional elementary facts for later use.

Proposition 4.5. *Let q be a horizontally valid move. If $\alpha \in [1, 2)$, then $\widehat{q}(\alpha, \beta) = 0$.*

Proof. We have

$$\begin{aligned} \widehat{q}(\alpha, \beta) &= \sum_y \left[\left(\sum_x q(x, y) \right) P_y(\beta) \right] && (35) \\ &= 0, && (36) \end{aligned}$$

as desired. □

Proposition 4.6. *Let q be a horizontally valid move. Then, $\widehat{q}(\alpha, 1) \geq \widehat{q}(\alpha, 2)$ for any $\alpha \in [1, \infty]$.*

Proof. By construction, $P_y(1) \geq P_y(2)$ for any y. We have

$$\widehat{q}(\alpha, 1) - \widehat{q}(\alpha, 2) = \sum_y \left[\left(\sum_x q(x, y) P_x(\alpha) \right) (P_y(1) - P_y(2)) \right]. \tag{37}$$

Both factors in the summand enclosed by brackets above are nonnegative, and thus the result follows. □

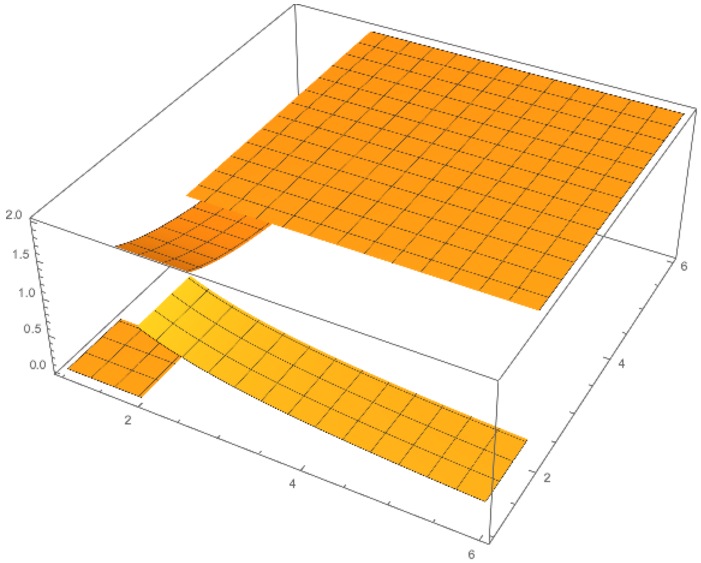


Figure 4: A graph of the function $\widehat{t_{1/10}}$. (Mathematica)

4.2 The target profiles

From subsection 3.3, we know that if there exists a quantum weak coin-flipping protocol whose bias is less than ϵ (with $0 < \epsilon < \frac{1}{2}$), then there must exist a valid TIPG (m_1, m_2) such that $m_1 + m_2$ is equal to

$$v_\epsilon \quad := \quad \left[\!\!\left[\frac{1}{2} + \epsilon, \frac{1}{2} + \epsilon\right]\!\!\right] - \frac{1}{2}[\![1, 0]\!] - \frac{1}{2}[\![0, 1]\!]. \tag{38}$$

We therefore have a crucial interest in the moves v_ϵ. However, it is mathematically simpler to instead study the family of moves

$$t_\tau \quad := \quad 2 \cdot [\![1, 1]\!] - [\![2 - \tau, 0]\!] - [\![0, 2 - \tau]\!] \tag{39}$$

for $0 < \tau < 1$. Note that if $\tau = 4\epsilon/(1 + 2\epsilon)$, then

$$v_\epsilon(x, y) \quad = \quad \frac{1}{2} t_\tau((2 - \tau)x, (2 - \tau)y), \tag{40}$$

and so (using Proposition 3.6), valid TIPGs for v_ϵ correspond exactly to valid TIPGs for t_τ via the same linear transformation.

If $R = (r_1, r_2)$ is a valid TIPG such that $r_1 + r_2 = t_\tau$, then we must have $\widehat{r_1} + \widehat{r_2} = \widehat{t_\tau}$. The profile function $\widehat{t_\tau}$ can be expressed as follows.

$$\widehat{t_\tau}(\alpha, \beta) \quad = \quad \begin{cases} 2 & \text{if } \alpha, \beta \geq 2 \\ 2 - P_{2-\tau}(\alpha) & \text{if } \alpha \geq 2, \beta < 2 \\ 2 - P_{2-\tau}(\beta) & \text{if } \alpha < 2, \beta \geq 2 \\ 0 & \text{if } \alpha, \beta < 2 \end{cases} \tag{41}$$

A graph of an example (with $\tau = 1/10$) is given in Figure 4.

5 Highly concentrated rational functions

This section adapts known complex analysis techniques to prove a result that will be needed in section 7.[2] I am grateful to Alexandre Eremenko for showing me the central method used in this section.

We will be concerned with rational functions on an interval $[a, b]$ that are highly concentrated — that is, rational functions that are significantly large at some interior point $c \in [a, b]$ and are more tightly bounded outside of a small neighborhood of c. Our interest (eventually) will be in studying deductions that we can make when such a function occurs in the profile of a move.

5.1 Preliminaries

We will first reproduce a standard result about the logarithm of the absolute value of an analytic function. We begin with the following observation: if f is an analytic function on a neighborhood of $\overline{\mathbb{D}}$ which has no zeroes in $\overline{\mathbb{D}}$, then there is a well-defined analytic function $\log f$ on a neighborhood of $\overline{\mathbb{D}}$ such that $2^{\log f} = f$. We have (see subsection 2.1):

$$\frac{1}{2\pi} \int_0^{2\pi} \log f(e^{it}) dt \quad = \quad \log f(0). \tag{42}$$

Since the real part of $\log f(z)$ is precisely $\log |f(z)|$, this proves the following.

Proposition 5.1. *Let f be an analytic function on an open neighborhood of $\overline{\mathbb{D}}$ such that f has no zeroes in $\overline{\mathbb{D}}$. Then,*

$$\log |f(0)| \quad = \quad \frac{1}{2\pi} \int_0^{2\pi} \log \left| f(e^{it}) \right| dt. \quad \square \tag{43}$$

For the result that we will prove in subsection 5.2, we will need a similar statement that addresses the case where f is permitted to have zeroes in $\overline{\mathbb{D}}$ and may not be analytic on $\mathbb{S}$. This motivates a somewhat more intricate claim. Let us say that a real-valued function on the unit circle $\mathbb{S}$ is a **step function** if it is locally constant at all but a finite number of points in $\mathbb{S}$. A proof of the following proposition is given in Appendix A.1.

Proposition 5.2. *Let f be a continuous function on $\overline{\mathbb{D}}$ which is analytic on $\mathbb{D}$. Suppose that $b \colon \mathbb{S} \to \mathbb{R}$ is a step function such that $\log |f(z)| \leq b(z)$ for any $z \in \mathbb{S}$. Then,*

$$\log |f(0)| \quad \leq \quad \frac{1}{2\pi} \int_0^{2\pi} b(e^{it}) dt. \quad \square \tag{44}$$

5.2 The complex values of a highly concentrated rational function

We will now prove the main result of this section. This result is concerned with rational functions that have real poles — that is, functions $f \colon \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ that can be expressed in the form

$$f(z) \quad = \quad \frac{g(z)}{\prod_{i=1}^n (z - c_i)}, \tag{45}$$

where $g(z)$ is a polynomial and $c_i \in \mathbb{R}$. We will show that under certain assumptions, these functions must take on large values on the complex unit circle.

Recall that if A and B are sets and $B \subseteq A$, then we write $A \smallsetminus B$ for the set of all elements in A that are not in B.

Proposition 5.3. *Let $f(z)$ be a rational function whose poles are all real and lie outside of $[-1, 1]$. Suppose that*

$$|f(0)| = 1, \tag{46}$$

and suppose that $\delta, \nu \in (0, 1)$ are such that

$$|f(z)| \leq \nu \qquad \textit{for all } z \in [-1, 1] \smallsetminus (-\delta, \delta). \tag{47}$$

[2] See sections III.1–III.2 in [21] for a more general discussion of some of the techniques used in the current section.

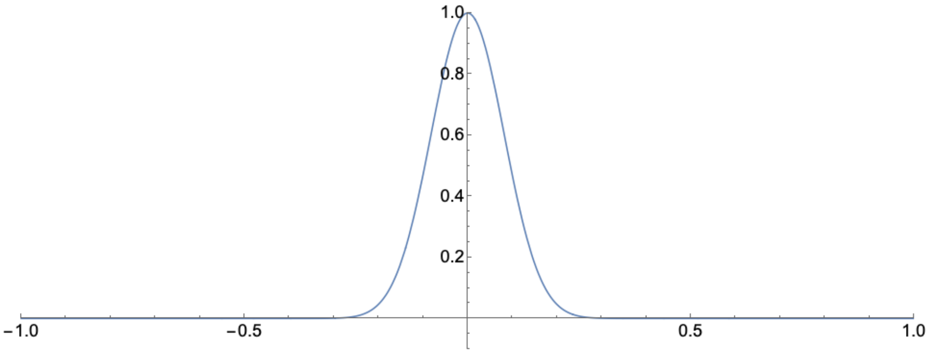


Figure 5: A graph over the interval $[-1, 1]$ of the example function in equation (49).

Then,

$$\max_{|z|=1} |f(z)| \quad \geq \quad \nu^{-\Omega(1/\delta)}. \tag{48}$$

We illustrate the statement above with an example. Let

$$h(z) \quad = \quad \left[\frac{4(z-1)(z+1)}{(z-2)(z+2)} \right]^{100}. \tag{49}$$

This function, which is graphed in Figure 5, satisfies $|h(0)| = 1$ and

$$|h(z)| \leq 0.1 \qquad \text{for all } z \in [-1,1] \smallsetminus (-0.2, 0.2). \tag{50}$$

The quantity $\max_{|z|=1} |h(z)|$ in this case is indeed quite large (on the order of 10^{20}).

Proof of Proposition 5.3. Our approach is to apply an analytic transformation which maps the unit circle $\mathbb{S}$ to the set

$$\mathbb{S} \cup ([-1,1] \smallsetminus (-\delta, \delta)) \tag{51}$$

and to thereby reduce the proof to an application of Proposition 5.2 above.

Let $G \colon \overline{\mathbb{D}} \to \overline{\mathbb{H}}$ be defined by

$$G(z) \quad = \quad i \cdot \frac{1+z}{1-z}. \tag{52}$$

Note that this function is a one-to-one mapping. Its inverse is given by

$$G^{-1}(z) \quad = \quad \frac{z-i}{z+i}. \tag{53}$$

Let $F \colon \overline{\mathbb{H}} \to \overline{\mathbb{H}}$ be the continuous function[3]

$$F(z) \quad = \quad \sqrt{\frac{z^2 |G(\delta)|^2 - 1}{|G(\delta)|^2 - z^2}}. \tag{54}$$

The function F maps ∞ to $G(\delta)$, maps 0 to $G(-\delta) = 1/G(\delta)$, and maps i to i. (Diagrams of F and G are given in Appendix A.2.) The image of $\mathbb{R} \cup \{\infty\}$ under F is the union of $\mathbb{R} \cup \{\infty\}$ together with the line segment from $G(\delta)$ to ∞, and the line segment from $G(-\delta)$ to 0.

[3] The rational function $z \mapsto \left(\frac{z^2 |G(\delta)|^2 - 1}{|G(\delta)|^2 - z^2} \right)$ on $\overline{\mathbb{H}}$ has two continuous square roots. We let F be the square root which maps $\overline{\mathbb{H}}$ into itself.

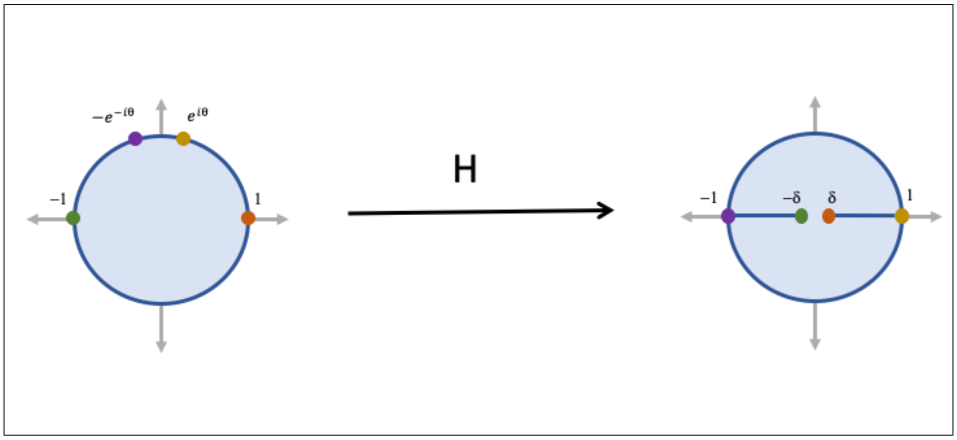


Figure 6: A diagram of the map H.

Now let H be defined by

$$H(z) = G^{-1}(F(G(z))). \tag{55}$$

Then,

$$\begin{aligned} H(0) &= 0 \qquad (56)\\ H(1) &= \delta \qquad (57)\\ H(-1) &= -\delta. \qquad (58)\end{aligned}$$

The image of the unit circle under H consists of the unit circle, the line segment from -1 to $-\delta$, and the line segment from δ to 1. Additionally, if we let $\theta \in [0, \pi/2]$ denote the angle of the unit-length complex number

$$\frac{i+\delta}{1+i\delta}, \tag{59}$$

then the following hold by direct computation:

$$\begin{aligned} H\left(e^{i\theta}\right) = H\left(e^{-i\theta}\right) &= 1 \qquad (60)\\ H\left(-e^{-i\theta}\right) = H\left(-e^{i\theta}\right) &= -1. \qquad (61)\end{aligned}$$

The points on the unit circle which lie clockwise between $e^{-i\theta}$ and $e^{i\theta}$ are mapped into the real interval $[\delta, 1]$, and the points which lie clockwise between $-e^{-i\theta}$ and $-e^{i\theta}$ are mapped into the real interval $[-1, -\delta]$. All other points on the unit circle remain on the unit circle under the application of H. A diagram of H is given in Figure 6.

We now compute upper bounds on the function $z \mapsto |f(H(z))|$. Let $M = \max_{|z|=1} |f(z)|$. Then, by our construction,

$$\begin{aligned} \log \left| f(H(e^{i\chi})) \right| &\leq \log M \qquad \text{if } \chi \in (\theta, \pi-\theta) \cup (\pi+\theta, 2\pi-\theta), \qquad (62)\\ \log \left| f(H(e^{i\chi})) \right| &\leq \log \nu \qquad \text{if } \chi \in [-\theta, \theta] \cup [\pi-\theta, \pi+\theta]. \qquad (63)\end{aligned}$$

Let $T\colon [0, 2\pi] \to \mathbb{R}$ be defined by

$$T(\chi) = \begin{cases} \log M & \text{if } \chi \in (\theta, \pi-\theta) \cup (\pi+\theta, 2\pi-\theta) \\ \log \nu & \text{otherwise.} \end{cases} \tag{64}$$

Then, applying Proposition 5.2,

$$\begin{aligned}\log|f(0)| &= \log|f(H(0))| && (65)\\ &\leq \frac{1}{2\pi}\int_0^{2\pi} T(s)ds && (66)\\ &= \left(1-\frac{2\theta}{\pi}\right)\log M + \left(\frac{2\theta}{\pi}\right)\log\nu. && (67)\end{aligned}$$

To complete the proof, it suffices to note that the quantity (59), which was used to define θ, is within distance $O(\delta)$ from the complex number i. Therefore, θ itself is within distance $O(\delta)$ from $\pi/2$. Thus we obtain

$$\log|f(0)| \leq O(\delta)(\log M) + (1-O(\delta))\log\nu. \qquad (68)$$

Since $\log|f(0)| = 0$ by assumption, we therefore have

$$\frac{-\log\nu}{O(\delta)} \leq \log M, \qquad (69)$$

which yields the desired result. □

We give a brief discussion to show how Proposition 5.3 can be useful for our purposes. Suppose that ℓ is a one-dimensional move such that

$$\widehat{\ell}(4) = 1 \qquad (70)$$

and that $\delta, \nu \in (0,1)$ are such that the inequality $\left|\widehat{\ell}(\alpha)\right| \leq \nu$ is always satisfied when

$$\alpha \in [3,5] \smallsetminus (4-\delta, 4+\delta). \qquad (71)$$

Then, for any $\alpha \in [3,5]$, the rational expression for $\widehat{\ell}(\alpha)$ is given by

$$\widehat{\ell}(\alpha) = \sum_x \ell(x)\cdot\frac{x\alpha - x}{x+\alpha-2}. \qquad (72)$$

By Proposition 5.3, there exists a unit-length complex number ζ such that

$$\left|\sum_x \ell(x)\cdot\left(\frac{x(4+\zeta)-x}{x+(4+\zeta)-2}\right)\right| \geq \nu^{-\Omega(1/\delta)}. \qquad (73)$$

Therefore,

$$\begin{aligned}\nu^{-\Omega(1/\delta)} &\leq \sum_x |\ell(x)|\cdot\frac{|x(4+\zeta)-x|}{|x+(4+\zeta)-2|} && (74)\\ &\leq \sum_x |\ell(x)|\cdot\frac{|x(4+1)-x|}{|x+(4-1)-2|} && (75)\\ &\leq \sum_x |\ell(x)|\cdot 4. && (76)\end{aligned}$$

Thus we conclude that $\|\ell\|_1 \geq v^{-\Omega(1/\delta)}$. Informally, this means that a one-dimensional move ℓ can only achieve a profile that is highly concentrated around $x=4$ if ℓ has exponentially large coefficients. This is similar to reasoning that we will use in section 7.

6 A lemma on random variables with bounded expectation

The following elementary lemma addresses a case where two constraints on the expectation of a random variable approximately determine the value of the random variable. This lemma will be used (in the context of some artificially constructed random variables) in order to carry out an intermediate step in the main proof of section 7.

Lemma 6.1. *There exists a universal function* $\mathbf{A}(u) \in O(\sqrt{u})$ *such that the following holds. If* X *is any positive real-valued random variable satisfying*

$$\begin{aligned} \mathbb{E}[X] &\leq 1, && (77) \\ \mathbb{E}[1/X] &\leq 1 + \delta, && (78) \end{aligned}$$

with $\delta > 0$*, then*

$$\mathbb{P}(|X - 1| < \mathbf{A}(\delta)) \quad \geq \quad \frac{2}{3}. \tag{79}$$

Proof. We have

$$\mathbb{E}[X + 1/X - 2] \quad \leq \quad \delta, \tag{80}$$

and $X + 1/X - 2 \geq 0$. Therefore,

$$\mathbb{P}[X + 1/X - 2 \geq 3\delta] \quad \leq \quad \frac{1}{3}, \tag{81}$$

or equivalently,

$$\mathbb{P}[X^2 - (2 + 3\delta)X + 1 < 0] \quad \geq \quad \frac{2}{3}. \tag{82}$$

By the quadratic formula, the event on the left side of inequality (82) is equivalent to

$$\left| X - (1 + \frac{3}{2}\delta) \right| \quad < \quad \sqrt{3\delta + \frac{9}{4}\delta^2}. \tag{83}$$

Thus with probability at least 2/3, $|X - 1|$ is less than

$$\frac{3}{2}\delta + \sqrt{3\delta + \frac{9}{4}\delta^2}. \tag{84}$$

The function above is in $O(\sqrt{\delta})$, and this completes the proof. □

7 The 1-norm of a time-independent point game

This section will perform most of the remaining technical work necessary to achieve our main result. Throughout this section, suppose that $\tau \in (0, 1)$, and that g is a horizontally valid move such that $g + g^\top = t_\tau$, where t_τ denotes the following move (see subsection 4.2):

$$t_\tau \quad = \quad 2[\![1, 1]\!] - [\![2 - \tau, 0]\!] - [\![0, 2 - \tau]\!]. \tag{85}$$

In this section we will show that the inequality $\|g\|_1 \geq \exp(\Omega(\tau^{-1/2}))$ must always hold. This is the result that will be used in section 8 to conclude that any weak coin-flipping protocol that achieves bias ϵ must involve at least $\exp(\Omega(\epsilon^{-1/2}))$ communication rounds.

For any $b \geq 0$, let $g_b \colon \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ denote the move defined by

$$g_b(x, y) \quad = \quad \begin{cases} g(x, y) & \text{if } y = b \\ 0 & \text{if } y \neq b. \end{cases} \tag{86}$$

(That is, g_b agrees with g on the horizontal line $y = b$, and is zero elsewhere.) Note that since g is horizontally valid, the profile function $\widehat{g_b}$ of g_b is always a nonnegative function.

7.1 Basic properties of g

By assumption, g is a horizontally valid move and its profile function $\widehat{g}$ satisfies

$$\widehat{g} + \widehat{g}^{\top} = \widehat{t_{\tau}}. \tag{87}$$

The function $\widehat{t_\tau}$ is written out explicitly in equations (41) and (22). The following facts are easily deduced.

Fact 7.1. *If $\alpha \in [2, \infty]$, then $\widehat{g}(\alpha, \alpha) = 1$.*

Fact 7.2. *If $\alpha, \beta \in [2, \infty]$, then $\widehat{g}(\alpha, \beta) \leq 2$.*

By Propositions 4.5 and 4.6, we have the following.

Fact 7.3. *If $\alpha \in [1, 2)$ and $\beta \in [1, \infty]$, then $\widehat{g}(\alpha, \beta) = 0$.*

Fact 7.4. *If $\alpha \in [2, \infty]$, then $\widehat{g}(\alpha, 1) = 2 - P_{2-\tau}(\alpha)$.*

Fact 7.5. *If $\alpha \in [2, \infty]$, then $\widehat{g}(\alpha, 2) \leq 2 - P_{2-\tau}(\alpha)$.*

It is easily seen from Definition 4.1 that for $\alpha \in [2, \infty]$,

$$\begin{aligned} P_{2-\tau}(\alpha) &\geq P_2(\alpha) - \tau \qquad (88)\\ &= \left(2 - \frac{2}{\alpha}\right) - \tau. \qquad (89)\end{aligned}$$

Thus from Fact 7.5 we have

Fact 7.6. *If $\alpha \in [2, \infty]$, then $\widehat{g}(\alpha, 2) \leq \frac{2}{\alpha} + \tau$.*

7.2 The isolating function

For any $a \in (2, \infty)$, we know (Fact 7.1) that

$$\sum_b \widehat{g_b}(a, a) = \widehat{g}(a, a) = 1, \tag{90}$$

and that each term $\widehat{g_b}(a, a)$ in the above summation is nonnegative. Our next goal is to show that the majority of the contribution to $\widehat{g}(a, a)$ comes from the terms $\widehat{g_b}(a, a)$ for which b is close to a. This is formally stated as follows.

Proposition 7.7. *There is a universal function $\mathbf{I}(u) \in O(\sqrt{u})$ such that for any $a \in [3, 5]$,*

$$\sum_{b : |b-a| < \mathbf{I}(\tau)} \widehat{g_b}(a, a) \quad \geq \quad \frac{2}{3}. \tag{91}$$

We note that the choice of the interval $[3, 5]$ in this statement is somewhat arbitrary — the proof method that follows can be used to prove the same statement over any interval $[p, q]$ for which $2 < p < q < \infty$ (with a different choice of function $\mathbf{I}$). The reason for making this type of restriction is that it facilitates calculations involving universal big-O error terms.

Proof of Proposition 7.7. Let Y be the stochastic map from $[3, 5]$ to $\mathbb{R}_{>0}$ defined by

$$\mathbb{P}(Y(a) = b) \quad = \quad \widehat{g_b}(a, a). \tag{92}$$

Note that by construction, for any $a \in [3, 5]$ and any $b > 0$, the function on $[1, \infty]$ defined by

$$\beta \quad \mapsto \quad \widehat{g_b}(a, \beta) \tag{93}$$

is a scalar multiple of the profile function of b (that is, $\beta \mapsto P_b(\beta)$). Therefore,

$$\widehat{g_b}(a, 2) = P_b(2) \cdot \frac{\widehat{g_b}(a, a)}{P_b(a)} = 1 \cdot \frac{\mathbb{P}(Y(a) = b)}{P_b(a)} \tag{94}$$

$$\widehat{g_b}(a, \infty) = P_b(\infty) \cdot \frac{\widehat{g_b}(a, a)}{P_b(a)} = b \cdot \frac{\mathbb{P}(Y(a) = b)}{P_b(a)} \tag{95}$$

which implies

$$\widehat{g}(a,2) = \mathbb{E}\left[\frac{1}{P_{Y(a)}(a)}\right] \tag{96}$$

$$\widehat{g}(a,\infty) = \mathbb{E}\left[\frac{Y(a)}{P_{Y(a)}(a)}\right], \tag{97}$$

where the expectations are taken over the random variable $Y(a)$. Expanding these expressions using the definition of the profile function, we have

$$\widehat{g}(a,2) = \mathbb{E}\left[\frac{Y(a)+a-2}{Y(a)\cdot a - Y(a)}\right] \tag{98}$$

$$= \mathbb{E}\left[\left(\frac{1}{a-1}\right) + \left(\frac{1}{Y(a)}\right)\left(\frac{a-2}{a-1}\right)\right] \tag{99}$$

and

$$\widehat{g}(a,\infty) = \mathbb{E}\left[Y(a)\left(\frac{1}{a-1}\right) + \left(\frac{a-2}{a-1}\right)\right]. \tag{100}$$

By Fact 7.6 and Fact 7.2,

$$\widehat{g}(a,2) \leq \frac{2}{a} + \tau, \tag{101}$$

$$\widehat{g}(a,\infty) \leq 2. \tag{102}$$

Combining the above two inequalities with formulas (99) and (100) above yields

$$\mathbb{E}\left[\left(\frac{1}{Y(a)}\right)\left(\frac{a-2}{a-1}\right)\right] \leq \frac{a-2}{a(a-1)} + \tau \tag{103}$$

$$\mathbb{E}\left[Y(a)\left(\frac{1}{a-1}\right)\right] \leq \frac{a}{a-1}. \tag{104}$$

Multiplying the equations above by $[a(a-1)/(a-2)]$ and $[(a-1)/a]$ respectively, we obtain

$$\mathbb{E}\left[\frac{a}{Y(a)}\right] \leq 1 + 7\tau \tag{105}$$

$$\mathbb{E}\left[\frac{Y(a)}{a}\right] \leq 1, \tag{106}$$

where, in (105), we used the fact that the function $a \mapsto [a(a-1)/(a-2)]$ on the interval $[3,5]$ does not exceed 7.

If we let $\mathbf{A}(u)$ be the function from Lemma 6.1, then

$$\mathbb{P}\left(\left|\frac{Y(a)}{a} - 1\right| < \mathbf{A}(7\tau)\right) \geq \frac{2}{3}. \tag{107}$$

Therefore,

$$\mathbb{P}\left(|Y(a) - a| < a\mathbf{A}(7\tau)\right) \geq \frac{2}{3}. \tag{108}$$

Letting $\mathbf{I}(u) := 5\mathbf{A}(7u)$ therefore yields

$$\mathbb{P}\left(|Y(a) - a| < \mathbf{I}(\tau)\right) \geq \frac{2}{3} \tag{109}$$

for any $a \in [3,5]$. By the definition of the stochastic map Y, this implies the desired result. □

We refer to $\mathbf{I}$ as the **isolating function**.

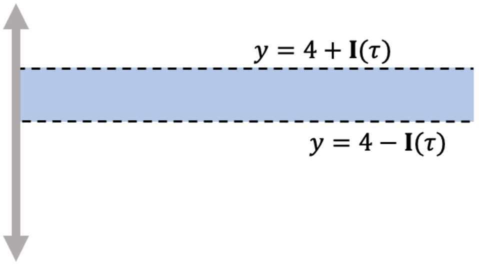


Figure 7: The move $\mathbf{g}$ is supported within the shaded blue region.

7.3 A lower bound on $\|g\|_1$

We are now ready to prove a lower bound on $\|g\|_1$ in terms of τ. We accomplish this by studying the behavior of the part of the move g that is concentrated near the horizontal line $y = 4$. Precisely, we will be concerned with the move

$$\mathbf{g} \quad := \quad \sum_{b : |b-4| < \mathbf{I}(\tau)} g_b, \tag{110}$$

where $\mathbf{I}$ is the isolating function from subsection 7.2. (See Figure 7.)

Proposition 7.8. *For all $a \in [3,5]$ such that $|a-4| \geq 2\mathbf{I}(\tau)$, we must have*

$$\widehat{\mathbf{g}}(a,a) \quad \leq \quad \frac{1}{3}. \tag{111}$$

Proof. For any such a, the move

$$\sum_{b : |b-a| < \mathbf{I}(\tau)} g_b, \tag{112}$$

has disjoint support from that of $\mathbf{g}$. The sum of the profile of (112) and the profile of $\mathbf{g}$ is therefore upper bounded by the profile of g. By Proposition 7.7 and Fact 7.1, we have $\widehat{\mathbf{g}}(a,a) \leq \frac{1}{3}$, as desired. □

Proposition 7.9. *The* 1*-norm $\|g\|_1$ of g satisfies*

$$\|g\|_1 \quad \geq \quad \exp(\Omega(\tau^{-1/2})). \tag{113}$$

Proof. Consider the function on $(2,\infty)$ given by

$$z \quad \mapsto \quad \widehat{\mathbf{g}}(z,z). \tag{114}$$

This function can be written out explicitly as follows:

$$z \quad \mapsto \quad \sum_{x,y} \mathbf{g}(x,y) \cdot \frac{xz - x}{x + z - 2} \cdot \frac{yz - y}{y + z - 2} \tag{115}$$

Let $D \colon \mathbb{C} \cup \{\infty\} \to \mathbb{C} \cup \{\infty\}$ denote the complex-valued rational function given by expression (115). By Propositions 7.7 and 7.8, the function D satisfies $D(4) \geq 2/3$, and $0 \leq D(z) \leq 1/3$ for all z in the set

$$[3,5] \smallsetminus (4 - 2\mathbf{I}(\tau), 4 + 2\mathbf{I}(\tau)). \tag{116}$$

Applying Proposition 5.3 (with appropriate affine transformations), we find that there exists a unit-length complex number ζ such that

$$\begin{aligned} |D(4+\zeta)| \quad &\geq \quad \exp(\Omega(1/\mathbf{I}(\tau)) && (117) \\ &\geq \quad \exp(\Omega(\tau^{-1/2})). && (118) \end{aligned}$$

Next we obtain an upper bound on $|D(4+\zeta)|$ in terms of $\|g\|_1$. We have

$$|D(4+\zeta)| \quad \leq \quad \sum_{x,y} |\mathbf{g}(x,y)| \cdot \left| \frac{x(4+\zeta)-x}{x+(4+\zeta)-2} \right| \cdot \left| \frac{y(4+\zeta)-y}{y+(4+\zeta)-2} \right|. \tag{119}$$

It is easy to see that the second and third factors in the summand above are each no more than 4. Therefore,

$$|D(4+\zeta)| \quad \leq \quad 16\,\|\mathbf{g}\|_1 \tag{120}$$

$$\leq \quad 16\,\|g\|_1\,. \tag{121}$$

Combining the above bound with inequality (118) above yields the desired result. □

8 Main result

We can now tie together the results of section 3 and section 7 to achieve our main result.

Proposition 8.1. *Let $\epsilon \in (0, \frac{1}{2})$, and suppose that $M = (m_1, m_2)$ is a valid time-independent point game such that*

$$m_1 + m_2 \quad = \quad \left[\!\!\left[\frac{1}{2}+\epsilon, \frac{1}{2}+\epsilon \right]\!\!\right] - \frac{1}{2}\cdot [\![1,0]\!] - \frac{1}{2}\cdot [\![0,1]\!]. \tag{122}$$

Then,

$$\|m_1\|_1 + \|m_2\|_1 \quad \geq \quad \exp(\Omega(\epsilon^{-1/2})). \tag{123}$$

Proof. Define moves m_1', m_2' by

$$m_i'(x,y) \quad = \quad 2m_i\left(x\left(\frac{1}{2}+\epsilon\right), y\left(\frac{1}{2}+\epsilon\right)\right). \tag{124}$$

Then, (m_1', m_2') is a valid TIPG which achieves the move $2[\![1,1]\!] - [\![2-\tau, 0]\!] - [\![0, 2-\tau]\!]$, where $\tau = 4\epsilon/(1+2\epsilon)$. Moreover, if we let $m' = [m_1' + (m_2')^\top]/2$, then $(m', (m')^\top)$ is a symmetric valid TIPG that achieves the same move. By Proposition 7.9,

$$\|m'\|_1 \quad \geq \quad \exp(\Omega(\tau^{-1/2})). \tag{125}$$

It is obvious that $\|m_1\|_1 + \|m_2\|_1 \geq \|m'\|_1$ and that $\tau \leq O(\epsilon)$. This completes the proof. □

Theorem 8.2. *Suppose that $\mathbf{C}$ is an n-round weak coin-flipping protocol with cheating probabilities $P_A^* \leq \frac{1}{2}+\epsilon$ and $P_B^* \leq \frac{1}{2}+\epsilon$. Then,*

$$n \quad \geq \quad \exp(\Omega(\epsilon^{-1/2})). \tag{126}$$

Proof. Combining Proposition 8.1 with Theorem 3.11, we find that for any $\delta > 0$,

$$n \quad \geq \quad \exp(\Omega((\epsilon+\delta)^{-1/2})). \tag{127}$$

Since the above inequality is true for any positive real number δ, the desired result follows. □

9 Further directions

A natural next step would be to compute an explicit function which would serve as a lower bound for n in Theorem 8.2. This is a matter of tracing through the steps of the proof, and should not be difficult. Explicit bounds on n will open the door to searching for quantum weak coin-flipping protocols that are optimized for the number of communication rounds (at a particular bias ϵ).

One can also try to lower bound the amount of quantum memory needed to achieve weak coin-flipping for a given bias. As discussed in [1], the quantum memory used by a protocol is related to the size of the support of its point games. Some of the same techniques used in this paper might be applicable to proving lower bounds on quantum memory size.

Can a related impossibility result be proved for strong quantum coin-flipping? A. Kitaev showed that any strong coin-flipping protocol must have bias at least $\frac{\sqrt{2}}{2} - \frac{1}{2} \approx 0.207$. Meanwhile, Chailloux and Kerenidis [9] proved, by building on Mochon's work on quantum weak coin flipping with vanishing bias [19], that strong coin-flipping is possible with bias arbitrarily close to $\frac{\sqrt{2}}{2} - \frac{1}{2}$. One could try to prove that strong coin-flipping with bias approaching $\frac{\sqrt{2}}{2} - \frac{1}{2}$ requires a large amount of communication.

Lastly, I will note that although we have found that the moves (39) that define quantum coin-flipping are exponentially hard to achieve by valid point games, my experience so far suggests this is a uniquely difficult family of moves. It may be worth exploring whether there are other simple classes of moves that can be more easily achieved by valid point games, and exploring whether such classes could have applications to positive results in two-party cryptography.

References

[1] Dorit Aharonov, Andre Chailloux, Maor Ganz, Iordanis Kerenidis, and Loick Magnin. A simpler proof of the existence of quantum weak coin flipping with arbitrarily small bias. *SIAM Journal on Computing*, 45(3):633–679, 2016.

[2] Dorit Aharonov, Amnon Ta-Shma, Umesh V. Vazirani, and Andrew C. Yao. Quantum bit escrow. In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, pages 705–714, New York, NY, USA, 2000. ACM.

[3] Lars V. Ahlfors. *Complex Analysis*. McGraw-Hill, 3rd edition, 1979.

[4] Andris Ambainis. A new protocol and lower bounds for quantum coin flipping. *Journal of Computer and System Sciences*, 68(2):398 – 416, 2004. Special Issue on STOC 2001.

[5] Atul Singh Arora, Jérémie Roland, and Stephan Weis. Quantum weak coin flipping. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pages 205–216, New York, NY, USA, 2019. ACM.

[6] Manuel Blum. Coin flipping by telephone a protocol for solving impossible problems. *SIGACT News*, 15(1):23–27, January 1983.

[7] Anne Broadbent and Christian Schaffner. Quantum cryptography beyond quantum key distribution. *Designs, Codes and Cryptography*, 78(1):351–382, Jan 2016.

[8] Harry Buhrman, Matthias Christandl, and Christian Schaffner. Complete insecurity of quantum protocols for classical two-party computation. *Phys. Rev. Lett.*, 109:160501, Oct 2012.

[9] André Chailloux and Iordanis Kerenidis. Physical limitations of quantum cryptographic primitives or optimal bounds for quantum coin flipping and bit commitment. *SIAM Journal on Computing*, 46(5):1647–1677, 2017.

[10] André Chailloux, Iordanis Kerenidis, and Jamie Sikora. Lower bounds for quantum oblivious transfer. *Quantum Info. Comput.*, 13(1-2):158–177, January 2013.

[11] Richard Cleve. Limits on the security of coin flips when half the processors are faulty. In *Proceedings of the Eighteenth Annual ACM Symposium on Theory of Computing*, STOC '86, pages 364–369, New York, NY, USA, 1986. ACM.

[12] Zeph Landau. A tensor network view of Kitaev's lower bound for strong coin flipping. Video, available at (`https://www.youtube.com/watch?v=1VYezqof_2Q`). 2014.

[13] Hoi-Kwong Lo. Insecurity of quantum secure computations. *Phys. Rev. A*, 56:1154–1162, Aug 1997.

[14] Hoi-Kwong Lo and H. F. Chau. Is quantum bit commitment really possible? *Phys. Rev. Lett.*, 78:3410–3413, Apr 1997.

[15] Hoi-Kwong Lo and H.F. Chau. Why quantum bit commitment and ideal quantum coin tossing are impossible. *Physica D: Nonlinear Phenomena*, 120(1):177 – 187, 1998. Proceedings of the Fourth Workshop on Physics and Consumption.

[16] Dominic Mayers. Unconditionally secure quantum bit commitment is impossible. *Phys. Rev. Lett.*, 78:3414–3417, Apr 1997.

[17] Carlos Mochon. Quantum weak coin-flipping with bias of 0.192. In *45th Annual IEEE Symposium on Foundations of Computer Science*, pages 2–11, Oct 2004.

[18] Carlos Mochon. Large family of quantum weak coin-flipping protocols. *Phys. Rev. A*, 72:022341, Aug 2005.

[19] Carlos Mochon. Quantum weak coin flipping with arbitrarily small bias. arXiv:0711.4114v1, 2007.

[20] Tal Moran, Moni Naor, and Gil Segev. An optimally fair coin toss. *Journal of Cryptology*, 29(3):491–513, Jul 2016.

[21] Rolf Nevanlinna. *Analytic Functions.* Springer-Verlag, 1970.

[22] R. W. Spekkens and T. Rudolph. Degrees of concealment and bindingness in quantum bit commitment protocols. *Phys. Rev. A*, 65:012310, Dec 2001.

[23] R. W. Spekkens and Terry Rudolph. Quantum protocol for cheat-sensitive weak coin flipping. *Phys. Rev. Lett.*, 89:227901, Nov 2002.

[24] Stephanie Wehner, David Elkouss, and Ronald Hanson. Quantum internet: A vision for the road ahead. *Science*, 362(6412), 2018.

A Appendices

A.1 Proof of Proposition 5.2

We state some elementary facts.

Fact A.1. *If f is an analytic function on an open neighborhood of a compact set $R \subseteq \mathbb{C}$, and f is not the zero function, then f has only a finite number of zeroes in R.*

Fact A.2. *If $f : \overline{\mathbb{D}} \to \mathbb{C}$ is a continuous function, then the family of functions on $\mathbb{S}$ given by*

$$z \mapsto f(cz), \tag{128}$$

for $c \in [0, 1]$, converges uniformly to $f_{|\mathbb{S}}$ as $c \to 1$.

We next prove modified versions of Proposition 5.1.

Proposition A.3. *Let f be an analytic function on an open neighborhood of $\overline{\mathbb{D}}$ such that f has no zeroes on the unit circle $\mathbb{S}$. Then,*

$$\log |f(0)| \quad \leq \quad \frac{1}{2\pi} \int_0^{2\pi} \log \left| f(e^{it}) \right| dt. \tag{129}$$

Proof. The function $f(z)$ can be written as $f(z) = f_1(z) f_2(z)$, where f_1 is a polynomial with roots in $\mathbb{D}$, and f_2 has no zeroes on $\overline{\mathbb{D}}$. By Proposition 5.1,

$$\log |f_2(0)| \quad = \quad \frac{1}{2\pi} \int_0^{2\pi} \log \left| f_2(e^{it}) \right| dt, \tag{130}$$

while direct computation shows that

$$\log |f_1(0)| \quad \leq \quad \frac{1}{2\pi} \int_0^{2\pi} \log \left| f_1(e^{it}) \right| dt. \tag{131}$$

Summing equation (130) and inequality (131) yields the desired result. □

By affine transformation, we have the following corollary.

Corollary A.4. *Let f be an analytic function on an open neighborhood of a closed disc $\overline{\mathbb{D}}(z, r)$ such that f has no zeroes on $\mathbb{S}(z, r)$. Then,*

$$\log |f(z)| \quad \leq \quad \frac{1}{2\pi} \int_0^{2\pi} \log \left| f(z + re^{it}) \right| dt. \quad \square \tag{132}$$

Now we will prove the desired result.

Proof of Proposition 5.2. If f is the zero function, then Proposition 5.2 is trivial, so we will assume that f is not identically zero. Take any $\delta > 0$. By Fact A.2 above, we can find $\epsilon > 0$ such that the following inequality holds on the annulus $\overline{\mathbb{D}} \smallsetminus \mathbb{D}(0, 1-\epsilon)$:

$$\left| f(z) - f\left(\frac{z}{|z|}\right) \right| \quad \leq \quad \delta. \tag{133}$$

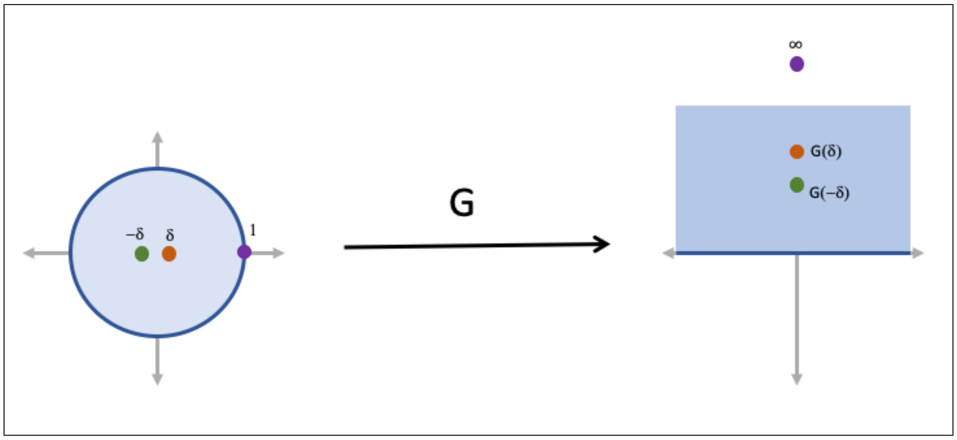
Choose (using Fact A.1) a real number $\zeta \in [\epsilon/2, \epsilon]$ such that f has no zeroes on $\mathbb{S}(0, 1-\zeta)$. Then, applying Corollary A.4,

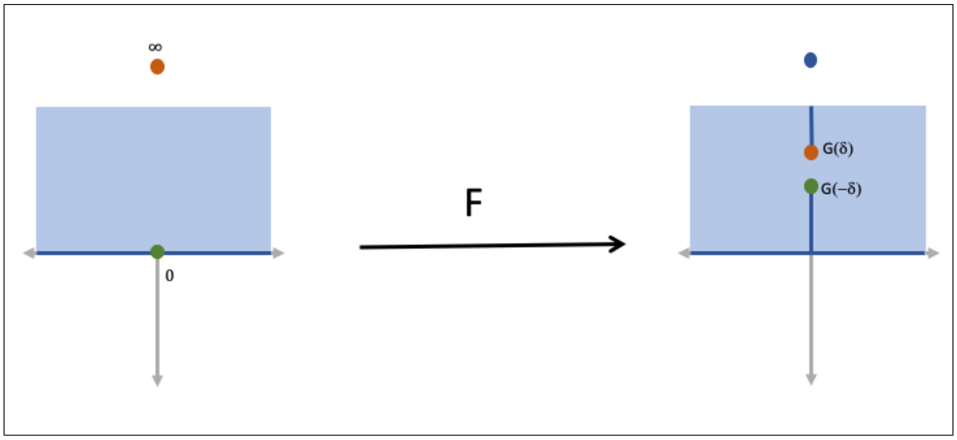
$$\begin{aligned} \log|f(0)| &\leq \frac{1}{2\pi}\int_0^{2\pi} \log\left| f((1-\zeta)e^{it})\right| dt & (134)\\ &\leq \frac{1}{2\pi}\int_0^{2\pi} \log\left(\left| f(e^{it})\right| + \delta\right) dt & (135)\\ &\leq \frac{1}{2\pi}\int_0^{2\pi} \log\left(2^{b(e^{it})} + \delta\right) dt. & (136)\end{aligned}$$

Since the upper bound (136) holds for any $\delta > 0$, the desired claim follows. □

A.2 Additional diagrams for subsection 5.2

Below, diagrams are given for the map G (52) and the map F (54).




∞
0
F
G(δ)
G(−δ)

Dynamic Spectrum Access with Reinforcement Learning for Unlicensed Access in 5G and Beyond

Susanna Mosleh[†] [§], Yao Ma[‡], Jacob D. Rezac[‡], and Jason B. Coder[‡]

[†]Associate, Communications Technology Laboratory, National Institute of Standards and Technology, USA
[§]Department of Physics, University of Colorado, Boulder, Colorado, USA
[‡]Communications Technology Laboratory, National Institute of Standards and Technology, USA

***Abstract*—Dynamic spectrum access (DSA) to achieve spectrum sharing in unlicensed bands is a promising approach for meeting the growing demands of forthcoming and deployed wireless networks, such as long-term evolution license-assisted access (LTE-LAA) and IEEE 802.11 Wi-Fi systems. In this paper, we consider a coexistence scenario where multiple LAA and Wi-Fi links compete for spectrum sharing subchannel access. We introduce a reinforcement-learning-based subchannel selection technique which allows access points (APs) and eNBs to select best subchannel distributively considering their medium access control (MAC) channel access protocols along with the physical layer parameters. The performance of this scheme is investigated through simulations, including the convergence property and sum throughput. Numerical results show that the proposed reinforcement-learning scheme converges fast and the sum throughput of the LAA and Wi-Fi systems is reasonably close to the result based on exhaustive search.**

***Index Terms*—5G and beyond, artificial intelligence, coexistence, LTE-LAA, MAC layer, PHY layer, Q-learning, WLAN.**

I. Introduction

Wireless communications are tightly integrated into our daily lives and promise to become more so in the future. Cisco Systems forecasts 3.6 mobile-connected devices per person by 2022 and a 7-fold increase in global mobile data traffic between 2017 and 2022 [1]. A promising approach to accommodating this growth is to transition from exclusively-licensed spectrum to a shared one, which offloads into unlicensed spectrum bands. One difficulty in this approach is balancing new network paradigms with incumbent networks, such as Wi-Fi.

Operating long term evolution (LTE) in unlicensed bands, such as with licensed assisted access (LAA), is a proposed solution to improve spectral-usage efficiency [2]. While such wireless coexistence proposals may have an enormous influence in solving spectrum usage issues, there are a number of challenges for both Wi-Fi and LTE networks to constructively share the spectrum. Assigning the spectrum in an intelligent and adaptive way may satisfy the diverse networks' requirements, build up an effective sharing of spectrum, and help to design future radio technologies (e.g., 5G NR-U [3]).

Innovations in artificial intelligence (AI) and machine learning (ML) are transformative in many industries. Wireless networking is also poised to benefit from these advances. For example, future 5G and beyond mobile devices are expected to access optimal spectral bands using highly-developed spectrum learning and inference. However, future networks will be immensely more intricate owing to complex network typologies and coordination schemes facilitating various end-user applications. Hence, it is computationally difficult for these future mobile devices to optimize key performance indicators (KPIs), such as throughput. Even more difficult to determine are strategies for efficient network management taking spectrum sharing into account. Nonetheless, as we demonstrate below, data-driven and adaptive machine learning algorithms are able to combat some of these difficulties to improve network performance.

Reinforcement learning (RL) is an area of ML and optimization which is well-suited to learning about dynamic and unknown environments [4]–[13]. It is especially suited to solving problems related to coexistence of wireless communications devices, as demonstrated by recent research. These works have been done mostly in the context of cognitive radio networks (CRNs). For example, [4] applied RL to estimate optimal channel access strategies for a single secondary user (SU) who dynamically selects one channel out of some number N channels. In [5], the authors assumed there is no primary user (PU) and applied RL to determine channel allocation for each SU without considering any collision with the PU. In [6], it was shown that the sensing capability, and therefore the transmission ability, could be enhanced for SUs with the proposed multi-agent RL (MARL)-based channel allocation. In [7], a MARL-based power control strategy was introduced to speed up the learning process of energy harvesting in communication systems.

A learning approach that accounts for the coexistence of LTE-U to model the resource allocation problem in LTE-U small stations (SBS), has been studied in [8], [9]. Specifically, to determine fair coexistence between LTE and Wi-Fi in the unlicensed spectrum, a Q-learning algorithm is applied in [8]. In [9] an RL algorithm based on long short-term memory (RL-LSTM) learning architecture was proposed to allocate the resources of LTE-U over the unlicensed spectrum. The most similar prior works ([10]–[13]) also used techniques from adaptive and reinforcement learning to study coexistence problems between Wi-Fi and other radio access technologies

Mosleh, Susanna; Ma, Yao; Rezac, Jake; Coder, Jason. "Dynamic Spectrum Access with Reinforcement Learning for Unlicensed Access in 5G and Beyond." Paper presented at 2020 IEEE 91st Vehicular Technology Conference: VTC2020-Spring, Antwerp, BE. May 25, 2020 - May 28, 2020.

(RATs). An adaptive MAC protocol for wireless sensor networks was proposed in [10], while the authors of [11] proposed a channel-selection method among LTE-U networks given only physical layer parameters. A learning method to select the duty cycle of LTE-U networks was presented in [12], [13].

The main contribution of our study is a novel reinforcement learning-based carrier-selection technique designed to ease the concurrent operation of Wi-Fi and LAA networks in unlicensed bands. Specifically, we use a Q-learning algorithm to select the spectrum in which a transmission node will operate based on self-learning experience. These transmission nodes each learn to maximize total network throughput without communicating with each other. Unlike earlier research on this topic [10]–[13], we take into account both MAC and physical layers and study the coexistence of two different types of network (LAA and Wi-Fi) operating simultaneously. To better reflect reality, we assumed there are different types of networks in which each base station serves several users. We consider the effects of collision between transmitting nodes. Moreover, in order to have a fully distributed strategy, there is neither a centralized controller in the network nor any information explicitly exchanged between different operators. The proposed algorithm enables both networks to constructively share the unlicensed spectrum and increase achieved data rates. It is worth noting that the introduced Q-learning algorithm could be applied to many types of communication systems, but LTE here is used as an example.

The remainder of this paper is organized as follows. Section II describes the system model and assumptions required for our analysis. Section III presents the problem formulation and introduces our proposed intelligent dynamic spectrum access. The impacts of the MAC and physical layer parameters in selecting a channel in a coexistence scenario is also explained in Section III. Simulation results are shown and discussed in Section IV. Finally, in Section V, an overview of the results and some concluding remarks are presented.

II. System Model

We consider a downlink coexistence scenario where two mobile network operators (MNOs) share for their operations the same unlicensed industrial, scientific, and medical (ISM) radio band. We are primarily interested in the operation of cellular base stations in an unlicensed band. However, the LTE base stations may have permission to utilize a licensed band as well. We assume each unlicensed band, indexed by $k \in \mathcal{K} \triangleq \{1, 2, \ldots, K\}$, can be shared between the MNOs in a time sharing fashion. The LAA network consists of n_L eNodeBs, while the Wi-Fi network is composed of n_W APs. The eNodeBs and APs are randomly distributed over a particular area, while LAA user equipment (UEs) and Wi-Fi clients/stations (STAs) are distributed around each eNodeB and AP, respectively, independently and uniformly. The transmission node $i \in \{\mathcal{L}, \mathcal{W}\}$, where $\mathcal{L} \triangleq \{n_\ell | \ell = 1, 2, \ldots, L\}$ and $\mathcal{W} \triangleq \{n_w | w = 1, 2, \ldots, W\}$, serves a set of $|\mathcal{U}_i|$ single antenna UEs/STAs on the unlicensed band, where $\mathcal{U}_i \triangleq \{u_{i,1}, u_{i,2}, \ldots, u_{i,|\mathcal{U}_i|}\}$. We assume that the transmission node i transmits with power p_i and the user association is based on the received power. Moreover, in order to meet the diverse network's requirement, we assume that eNodeBs may belong to different LTE operators with different priority classes (PCs) as introduced by the 3GPP standards committee [14]. Similarly, APs may belong to different Wi-Fi operators with four different access categories (ACs) [15]. We also assume *(i)* both Wi-Fi and LAA are in the saturated traffic condition, *(ii)* there is no hidden node problem[1] in the network (*i.e.,* every transmission node i is able to hear one another), and *(iii)*, the channel knowledge is ideal, so, the only source of packet failure (unsuccessful transmission) is collision.

Medium access in Wi-Fi is based on contention with random back-off. This process is known as carrier sense multiple access with collision avoidance (CSMA/CA) [16], [17]. In order to discover whether the channel is idle or busy, the station that accesses the medium should sense the channel by performing clear channel assessment (CCA). The distributed coordination function (DCF) operation progresses if the channel is found out to be idle. Otherwise, the transmitting station refrains from transmitting data until it senses the channel is available. Similarly, LTE-LAA uses a listen before talk (LBT) channel access mechanism to maintain fair coexistence with the Wi-Fi. Among different LAA-LBT schemes, Cat 4 LBT, which is based on the Wi-Fi CSMA/CA scheme, is well-suited to coexistence [2]. Although LTE and Wi-Fi technologies follow the same channel access procedure, they select different carrier sense mechanisms, different channel sensing threshold levels, and different channel contention parameters, leading to different unlicensed channel access probabilities and thus, different throughput.

In this section, we will briefly derive the normalized network throughput of both systems on each unlicensed band, a quantity we aim to optimize below. Conforming with the analytical model in [18], the probability of packet transmission by a transmitting node i in a randomly-chosen time slot on the k-th unlicensed channel can be written as

$$p_{\text{tr},i}^{(k)} = \frac{2(1-2p_{c,i}^{(k)})}{(1-2p_{c,i}^{(k)})(1+\text{CW}_{\min,i}^{(k)})+p_{c,i}^{(k)}\text{CW}_{\min,i}^{(k)}(1-(2p_{c,i}^{(k)})^{m_i^{(k)}})}, \quad (1)$$

where $\text{CW}_{\min,i}^{(k)}$ and $m_i^{(k)}$ are the minimum contention window size and the maximum back-off stage of the transmitting node i on the unlicensed band k, respectively, and $p_{c,i}^{(k)}$ is the probability of collision experienced by the i-th transmitting node on the k-th unlicensed channel. The probability of collision experienced by n_w AP and n_ℓ eNodeB can be expressed as

$$\begin{aligned} p_{c,n_w}^{(k)} &= 1 - \prod\nolimits_{\acute{w} \neq w} (1 - p_{\text{tr},n_{\acute{w}}}^{(k)}) \prod\nolimits_{\ell} (1 - p_{\text{tr},n_\ell}^{(k)}), \\ p_{c,n_\ell}^{(k)} &= 1 - \prod\nolimits_{w} (1 - p_{\text{tr},n_w}^{(k)}) \prod\nolimits_{\acute{\ell} \neq \ell} (1 - p_{\text{tr},n_{\acute{\ell}}}^{(k)}), \qquad (2) \end{aligned}$$

respectively, where $\acute{w} = 1, \ldots, W$ and $\acute{\ell} = 1, \ldots, L$.

[1] Here, we assume perfect spectrum sensing in both systems. Therefore, there are neither hidden nodes nor false alarm/miss detection problems in the network. The impact of imperfect sensing is beyond the scope of this paper and investigating the effect of CCA errors is an important topic for future work.

The probability of collision can be split into three parts: the probability of collision due to the collision between the Wi-Fi transmissions, between the LAA transmissions, and between the Wi-Fi and the LAA transmissions, respectively given by

$$p^{(k)}_{c,\mathcal{W}} = (1 - p^{(k)}_{tr,\mathcal{L}})\big[p^{(k)}_{tr,\mathcal{W}} - \sum\nolimits_{w} p^{(k)}_{\text{tr},n_w} \prod\nolimits_{\acute{w}\neq w}(1 - p^{(k)}_{\text{tr},n_{\acute{w}}})\big],$$
$$p^{(k)}_{c,\mathcal{L}} = (1 - p^{(k)}_{tr,\mathcal{W}})\big[p^{(k)}_{tr,\mathcal{L}} - \sum\nolimits_{\ell} p^{(k)}_{\text{tr},n_\ell} \prod\nolimits_{\acute{\ell}\neq \ell}(1 - p^{(k)}_{\text{tr},n_{\acute{\ell}}})\big],$$

and $p^{(k)}_{c,\mathcal{W},\mathcal{L}} = p^{(k)}_{tr,\mathcal{L}} \cdot p^{(k)}_{tr,\mathcal{W}}$. Here $p^{(k)}_{tr,\mathcal{L}} = 1 - \prod_{\ell=1}^{L}(1 - p^{(k)}_{\text{tr},n_\ell})$ and $p^{(k)}_{tr,\mathcal{W}} = 1 - \prod_{w=1}^{W}(1 - p^{(k)}_{\text{tr},n_w})$ denote the LAA's and Wi-Fi's probability of transmission on the k-th unlicensed channel, respectively. Moreover, the probability of a successful transmission by the i-th transmitter on the k-th unlicensed band can be written as

$$p^{(k)}_{s,n_w} = p^{(k)}_{\text{tr},n_w} \prod\nolimits_{\acute{w}\neq w}(1 - p^{(k)}_{\text{tr},\acute{w}}) \prod\nolimits_{\ell}(1 - p^{(k)}_{\text{tr},n_\ell}),$$
$$p^{(k)}_{s,n_\ell} = p^{(k)}_{\text{tr},n_\ell} \prod\nolimits_{\acute{\ell}\neq \ell}(1 - p^{(k)}_{\text{tr},\acute{\ell}}) \prod\nolimits_{w}(1 - p^{(k)}_{\text{tr},n_w}). \quad (3)$$

Hence, the average length of a time slot in the unlicensed channel k can be calculated as

$$T^{(k)}_{\text{avg}} = (1 - p^{(k)}_{tr})\mathbb{E}\{T^{(k)}_{\text{idle}}\} + p^{(k)}_{s,\mathcal{W}}\mathbb{E}\{T^{(k)}_{s,\mathcal{W}}\} + p^{(k)}_{s,\mathcal{L}}\mathbb{E}\{T^{(k)}_{s,\mathcal{L}}\}$$
$$+ p^{(k)}_{c,\mathcal{W}}\mathbb{E}\{T^{(k)}_{c,\mathcal{W}}\} + p^{(k)}_{c,\mathcal{L}}\mathbb{E}\{T^{(k)}_{c,\mathcal{L}}\} + p^{(k)}_{c,\mathcal{W},\mathcal{L}}\mathbb{E}\{T^{(k)}_{c,\mathcal{W},\mathcal{L}}\},$$

where $p^{(k)}_{tr}$ is the probability of occupation of the k-th unlicensed channel and can be expressed as

$$p^{(k)}_{tr} = 1 - \prod\nolimits_{w=1}^{W}(1 - p^{(k)}_{\text{tr},n_w}) \prod\nolimits_{\ell=1}^{L}(1 - p^{(k)}_{\text{tr},n_\ell}),$$

and $p^{(k)}_{s,\mathcal{W}} = \sum_{w} p^{(k)}_{s,n_w}$ and $p^{(k)}_{s,\mathcal{L}} = \sum_{\ell} p^{(k)}_{s,n_\ell}$ denote the successful transmission probability of the whole Wi-Fi and LAA networks on the k-th unlicensed band, respectively. Moreover, $T^{(k)}_{s,\mathcal{L}}$, $T^{(k)}_{s,\mathcal{W}}$, $T^{(k)}_{c,\mathcal{W}}$, $T^{(k)}_{c,\mathcal{L}}$, and $T^{(k)}_{c,\mathcal{W},\mathcal{L}}$ indicate the time that the k-th channel is occupied by an LAA successful transmission, a Wi-Fi successful transmission, a collision among the Wi-Fi transmissions, a collision among the LAA transmissions, and a collision between the Wi-Fi and the LAA transmissions, respectively. Considering the basic access scheme into account [18], [19], $T^{(k)}_{s,\mathcal{W}} = T_{P,\mathcal{W}} + T_{\text{SIFS}} + T^{(k)}_{\text{idle}} + T_{\text{ACK}} + T_{\text{DIFS}} + T^{(k)}_{\text{idle}} + (\text{PHY}_{\text{header}} + \text{MAC}_{\text{header}})/R^{(k)}_{\mathcal{W}}$, $T^{(k)}_{c,\mathcal{W}} = T_{P,\mathcal{W}} + (\text{PHY}_{\text{header}} + \text{MAC}_{\text{header}})/R^{(k)}_{\mathcal{W}} + T_{\text{DIFS}} + T^{(k)}_{\text{idle}}$, $T^{(k)}_{s,\mathcal{L}} = T_{P,\mathcal{L}} + T_{\text{SIFS}} + T_{\text{ACK}} + T_{\text{DIFS}} + T^{(k)}_{\text{idle}}$, $T^{(k)}_{c,\mathcal{L}} = T^{(k)}_{s,\mathcal{L}}$, and $T^{(k)}_{c,\mathcal{W},\mathcal{L}} = \max(T^{(k)}_{c,\mathcal{L}}, T^{(k)}_{c,\mathcal{W}})$; where $T_{P,\mathcal{W}}$ ($T_{P,\mathcal{L}}$) indicates the Wi-Fi (LAA) payload duration, $\text{PHY}_{\text{header}}+\text{MAC}_{\text{header}}$ presents the packet header, and T_{SIFS}, T_{ACK}, and T_{DIFS} refer to short interframe space (SIFS), acknowledgement signal duration, and distributed interframe space (DIFS), respectively.

Finally, the normalized network throughput of LAA and Wi-Fi systems on the k-th unlicensed band as a function of both MAC and Physical layers parameters and can be expressed as

$$S^{(k)}_{\mathcal{L}} = p^{(k)}_{s,\mathcal{L}} T_{P,\mathcal{L}} R^{(k)}_{\mathcal{L}} / T^{(k)}_{\text{avg},\mathcal{L}}, \quad (4)$$
$$S^{(k)}_{\mathcal{W}} = p^{(k)}_{s,\mathcal{W}} T_{P,\mathcal{W}} R^{(k)}_{\mathcal{W}} / T^{(k)}_{\text{avg},\mathcal{W}},$$

where $T^{(k)}_{\text{avg},\mathcal{W}}$ ($T^{(k)}_{\text{avg},\mathcal{L}}$) denotes the average time duration to achieve a successful transmission in the Wi-Fi (LAA) network and $R^{(k)}_{\mathcal{W}}$ ($R^{(k)}_{\mathcal{L}}$) refers to the Wi-Fi's (LAA's) physical data rate, which can be calculated as follows. Assuming each channel k can be shared between the UEs associated with the eNodeB n_ℓ in a time sharing fashion, the physical data rate of the LAA network on the k-th unlicensed band can be expressed as [20]

$$R^{(k)}_{\mathcal{L}} = \sum\nolimits_{\ell=1}^{L} \sum\nolimits_{i=1}^{|\mathcal{U}_{n_\ell}|} c^{(k)}_{u_{n_\ell},i} B_k \log_2(1 + \chi^{(k)}_{u_{n_\ell},i} \cdot \text{SINR}^{(k)}_{u_{n_\ell},i}),$$

where $0 \leq c^{(k)}_{u_{n_\ell},i} \leq 1$ denotes the time sharing component of the UE $u_{n_\ell,i}$ on the k-th unlicensed band that fulfills $0 \leq \sum_{n_\ell \in \mathcal{N}_L} \sum_{u_{n_\ell},i \in \mathcal{U}_{n_\ell}} c^{(k)}_{u_{n_\ell},i} \leq 1$, B_k is the bandwidth of the k-th channel, and $\chi^{(k)}_{u_{n_\ell},i}$ is the channel access indicator which is 1 if the eNodeB n_ℓ serves the UE $u_{n_\ell,i}$ on the k-th unlicensed channel, and is 0 otherwise. Moreover, $\text{SINR}^{(k)}_{u_{n_\ell},i}$ represents the signal-to-interference-plus-noise ratio (SINR) for serving user $u_{n_\ell,i}$ on the k-th unlicensed band. Following the same notation, the physical data rate of the Wi-Fi's network can be written as [20]

$$R^{(k)}_{\mathcal{W}} = \sum\nolimits_{w=1}^{W} \sum\nolimits_{j=1}^{|\mathcal{U}_{n_w}|} B_k \log_2(1 + \chi^{(k)}_{u_{n_w},j} \cdot \text{SINR}^{(k)}_{u_{n_w},j}).$$

It is worth noting that in the case where an eNodeB/AP aggregates multiple unlicensed bands, the total throughput may be obtained by summation over all unlicensed channels.

III. Intelligent Dynamic Spectrum Access

If both Wi-Fi and LAA on the unlicensed band select channels appropriately, the spectral efficiency of the whole network will improve. Equation (4) reveals the impact of channel selection by each MNO on network's throughput. Specifically, if transmitter i selects channel k and channel k is not occupied by the other transmission nodes, then the probability of successful transmission by the i-th node in the k-th channel increases, leading to higher throughput. Moreover, if each channel is selected such that interference is avoided (or at least minimized) among the transmission nodes, there will be a higher SINR, leading to a higher physical data rate and throughput. In this section, we employ an algorithm for channel selection that learns from previous experience to improve network throughput. The learning algorithm also makes adaptive use of channel usage information.

We introduce a Q-Learning based dynamic channel selection algorithm here for both MNOs. The aim is to enhance the throughput of both Wi-Fi and LAA networks on the unlicensed band. Q-Learning is a value-based reinforcement learning technique that discovers the best action at each state by maximizing a value function, typically denoted by Q, that returns the expected reward of each action-state pair [21]. Here, "actions" are ways to explore the environment, and "states" describe the environment (we give an example for the coexistence problem below). The goal is to learn how to map environmental conditions into actions that maximize a reward. By interacting with the environment and trying different actions over time, the algorithm learns which action (in each state) provides the highest reward. Employing the Q-learning algorithm in a coexistence scenario allows each transmission node to learn to select the channel that yields the best throughput based on its interaction with the environment.

To be specific, we consider a set of $\mathcal{L}$ eNodeBs and $\mathcal{W}$ APs as the transmitters which will take actions. Denote the ith transmitter's set of actions by $\mathcal{A}_i = \{a_{i,1}, \ldots, a_{i,|\mathcal{A}_i|}\}$ and states $\mathcal{S}_i = \{s_{i,1}, \ldots, a_{i,|\mathcal{S}_i|}\}$. There is an associated set of Q-values related to each state-action pair. This is often called a Q-table, and it is unknown a priori. In the proposed algorithm, we consider four states for each channel observed by each transmission node: Idle, Successful transmission, Collision, and Contention. In the intelligent dynamic spectrum access of our interests, a collision is detected if both transmitters select the same channel for their operations, both have at least one user to serve, and the energy that they received from each other is less than predefined energy detectors threshold. The observed state is Contention if the initial clear channel assessment returns busy, either immediately or during the DIFS, or if the transmitter wants to contend after a successful transmission. The actions correspond to picking a different transmission band out of $|\mathcal{K}|$ bands in $\mathcal{K}$ and $|\mathcal{K}| = |\mathcal{A}_i|$.

The Q-table determines the best action at each state as judged by the action-state pair which yields highest reward. Particularly, the Q-value at each action-state pair is initialized with an estimate of the future reward. That results from a transmitter selecting a particular action while it is in a certain state. Then, based on interaction and feedback with the environment (e.g., MAC layer parameters, channel propagation model, topology of the Wi-Fi and LAA networks, sensing threshold), the transmitter learns the outcome of that action-state pair. In the proposed algorithm, each transmission node i measures $Q(s_{i,j}, a_{i,k})$, the expected reward achieved by selecting the k-th channel when the i-th transmission node is in the j-th state. If the k-th channel has never been selected by the transmission node i in the past, then $Q(s_{i,j}, a_{i,k})$ value is fixed to an arbitrary value Q_0 at the initialization step.

We denote by $R(s_{i,j}, a_{i,k}) := S_{\mathcal{L}}^{(k)} + S_{\mathcal{W}}^{(k)}$ the reward resulting from environmental interaction. Since our goal is to select the channel that maximizes the spectral efficiency of the whole network, we define this reward as the average throughput acquired by the i-th transmission node when it is in the j-th state and selects the k-th action (channel). The Q-value is then iteratively updated by

$$Q_{t+1}(s_{i,j}, a_{i,k}) \leftarrow (1-\alpha)Q_t(s_{i,j}, a_{i,k}) + \alpha R_{t+1}(s_{i,j}, a_{i,k})$$

where $0 \leq \alpha \leq 1$ is the learning rate and determines how fast the learning happens [21]. As is typically the case in learning algorithms, selecting α requires some care. Indeed, a small value of α results in a long learning process (which could be detrimental in practice), while a large value of α could cause the algorithm to not converge. As soon as the Q-value of the current state-action is updated according to the above equation, the i-th transmission node selects a channel corresponding to its new state.

Note that in order to find the effect of each selected channel, the transmission node needs to explore the environment. This suggests that the transmission nodes not only select the channels that provide the current highest estimated throughput, but also take on new actions to find better channel selection strategies. There are many methods for choosing action-selection strategies to balance this exploration of the environment with reward maximization. Examples are greedy action-selection, ϵ-greedy action-selection, and softmax action-selection [21]. In the greedy action-selection strategy, the transmission node always picks the channel that provides the highest Q-value and never explores any new channels. Alternatively, in the ϵ-greedy action-selection method, the transmission node uniformly picks a channel with probability ϵ. Although this strategy attempts to balance the exploration of the environment with maximizing reward, the chance of selecting the worst-appearing channel is the same as the chance of picking any other channel, due to the uniform selection probability. The softmax action-selection policy, like the above aforementioned action-selection strategies, gives the highest selection probability to the greedy action. In contrast, however, the rest of the actions are weighted according to their estimated values. Under this policy, the probability of picking the k-th unlicensed channel by the i-th transmission node is

$$Pr(i,k) = \exp\left(\frac{Q(s_{i,j}, a_{i,k})}{\tau(i)}\right) \Big/ \sum_{m=1}^{K} \exp\left(\frac{Q(s_{i,j}, a_{i,m})}{\tau(i)}\right),$$

where $\tau(i) > 0$ is the so-called temperature. With a high value of $\tau(i)$, the softmax policy is nearly identical to the ϵ-greedy action-selection strategy and so, the probability of selecting the different channels is almost the same. In contrast, selecting $\tau(i)$ with a low value makes a big difference in selection probability of an action that varies in Q-value. As a result, the probability of selecting a channel increases when it has a larger Q-value. Then, as $\tau(i)$ goes to zero, the softmax policy becomes the greedy action-selection strategy. Hence, we select the softmax policy to best balance exploration and reward maximization. Moreover, as a means to decrease the amount of exploration by the transmission node i (when the number of channel selection by this node increases), we consider a cooling function [22] that can be expressed as

$$\tau(i) = \tau_0 / \log_2(1 + n(i)),$$

where τ_0 denotes the initial temperature and $n(i)$ represents the number of channels that have been selected by the i-th transmission nodes.

Remark 1. *Compared to existing approaches, Q-learning is at an advantage. Indeed, the Q-learning approach is model-free and becomes adjusted to the new environment changes quickly. Therefore, there is no need to have the full knowledge of state transition in a dynamic environment. In addition, and in terms of computational complexity, Q-learning has much less complexity compared to the tradition optimization approaches such as branch and bound algorithm. To be specific, Q-learning is computationally limited by the size of the action and state spaces. Q-learning is quadratic in the number of st-*

ates and linear in the number of actions, e.g., $\mathcal{O}(|\mathcal{S}|^2|\mathcal{A}|)$.

IV. Simulation Results and Analysis

We evaluate the performance of the proposed algorithm in a coexistence scenario. We simulate a scenario in which 3 eNodeBs compete for the unlicensed channels with 3 APs. All transmitters are randomly distributed over an area of size 120×80 m^2 with minimum distance 40 meters, as shown in Fig. 1. All UEs and Wi-Fi clients are independently and uniformly distributed around each eNodeB and AP, respectively. We consider 5 UEs (Wi-Fi clients) per eNodeB (AP). Each UE (Wi-Fi client) is assigned to the eNodeB (AP) that provides it with the highest received power. The antenna height of the transmission nodes and users are 6 meters and 1.5 meters, respectively. The carrier frequency is 5 GHz and each channel bandwidth is 20 MHz. The path-loss and shadowing between transmission nodes, and between transmission nodes and users are generated following [23] for the indoor scenario. The transmit power at each transmission node is fixed to 23 dBm while the noise figure and the thermal noise level at each user is set to 9 dB and -174 dBm/Hz, respectively. Moreover, we assume the omni-directional antenna pattern with a 0 dBi antenna gain. While Wi-Fi uses the CSMA/CA scheme to access the unlicensed band with CCA-CS and CCA-ED thresholds of -82 dBm and -62 dBm, respectively [15], an LBT with random back-off and variable contention window size with CCA-ED threshold of -72 dBm is considered for the LAA channel access scheme [14]. The MAC layer parameters are given in Table I, and the initial Q-learning parameters are set to $\alpha = 0.1$, $Q_0 = 0.5$, and $\tau_0 = 0.15$.

According to this geometry and propagation model, only some pairs of transmitters can detect each other. We summarize which transmitters can detect each other in matrix $\mathbf{I}_D$ given below, in which the ij-th element of $\mathbf{I}_D$ is nonzero only if transmission node i detects transmission j. In the case of detection, this element is 1. The first three rows (columns) corresponds to AP1 to AP3 while eNodeB1 to eNodeB3 are organized in the last three rows (columns). For the aforementioned scenario the matrix $\mathbf{I}_D$ is given as

$$\mathbf{I}_D = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5}$$

Note that in calculating the reward function and updating the Q-table in the proposed algorithm, we need to determine the throughput. For computing the throughput, we need to determine SINR. The $\mathbf{I}_D$ matrix, which tells us which pairs of transmitters can detect each other, helps us to calculate the interference and hence the SINR, and throughput.

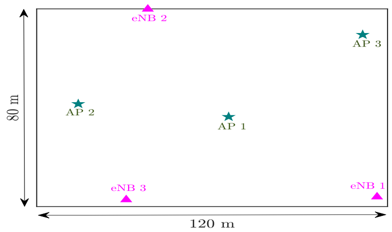


Fig. 1: Simulation layout

TABLE I: MAC Layer Parameters

Parameter	value
LAA's packet payload duration	1 ms
Wi-Fi's packet payload duration	1 ms
MAC header	272 bits
PHY header	128 bits
ACK	112 bits + PHY header
SIFS	16 μs
DIFS	34 μs
Idle slot time	9 μs
Wi-Fi contention window size	16
LAA contention window size	16
Wi-Fi maximum backoff stage	6
LAA maximum backoff stage	3

We consider the case of 6 different available channels in which all active transmission nodes in the area apply the proposed Q-learning method. Neither Wi-Fi nor LAA has knowledge about the other operators' selected channels. Each node *(i)* measures the expected reward achieved by selecting each channel when it is in a specific state, *(ii)* updates its Q-table, and *(iii)* chooses the channel that maximizes the whole network throughput.

The channel selection probability of each transmission node is shown in Fig. 2. At the initialization step, each transmission node interacts with the environment to explore different channels. It is worth mentioning that if the radio conditions change, the transmitters keep interacting with the environment until they determine which channel best improves the whole network throughput. With a change in radio conditions we will see that a learner must re-learn its policies as well. It is expected that any changes in the radio conditions during the initialization step will increase the initialization step duration. When the initialization step is done, according to the feedback from the environment and past experience, each node selects a channel that maximizes the whole network throughput, with a high probability. Since AP1 detects all other transmission nodes, it selects a fully vacant channel. AP2 and AP3 cannot detect each other, and they select the same channel. eNodeB1 only detects AP1. At first eNodeB1 selects the 4-th channel for its transmission, but when AP1 selects this channel, the eNodeB1 learns that this channel is now occupied, and switches to channel 2. The eNodeB2 only detects AP1 and AP2, so learns that the channels occupied by these two nodes do not maximize its throughput, and therefore selects a different channel. However, this channel is using by eNodeB1 that eNodeB2 does not detect. Finally, eNodeB3 selects the first channel which none of the other nodes picks as the means of maximizing the whole network spectral efficiency.

Simulation time is measured with respect to algorithmic time steps. Each time step is equal to the transmission nodes payload duration, i.e., 1 ms. Taking the MAC layer parameters

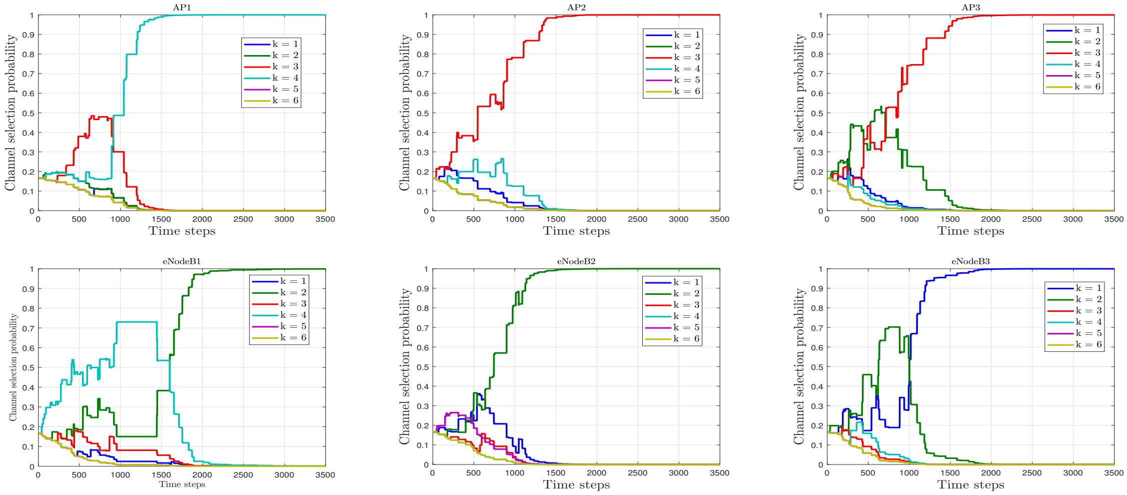


Fig. 2: Channel selection probability of each transmission node versus the time steps.

into account, all the APs and eNodeBs generate activity periods where they transmit data to their users with an average of 69- and 77-time steps. Since, the duration of initial time to learn the best solution is in order of 800-time steps in our simulations, the decisions of selecting a channel for transmitting the data are made after trying 10 initial channel selections.

It is worth mentioning that our simulation did not converge to a simple solution of allocating each transmitter to its own channel, instead favoring a scenario in which the transmitters choose the same channels. The reason for this solution is related to the physical scenario in which we pose the problem, as well as the MAC and physical layer parameters used by the APs and eNodeBs.

In Fig. 3 we consider a challenging case in which there are only 4 channels available. All active transmission nodes apply the proposed Q-learning method. For this scenario, again, we randomly distribute all transmitters and receivers over the area as discussed earlier. According to this geometry and propagation model the new matrix $\mathbf{I}_D$ is given as follows

$$\mathbf{I}_D = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \tag{6}$$

Since there are only 4 available channels, sharing the same spectrum bands in the time domain is inescapable. Fig. 3 represents the channel selection probability of each transmission node. As it is shown in Fig. 3 after an initialization step that all channels are evaluated, AP 2 and eNodeB 2 use the same channel. AP 2 and eNodeB 2 are located far from each other and they are not reciprocally detected during the sensing period. Hence, they can utilize the same channel without sharing it in the time domain. Note that if these two transmitters detected each other and decided to select a same channel, they would share the channel in the time domain obeying the LBT requirements. The same discussion is applied to AP3 and eNodeB 1. AP1 and eNodeB 3 learn to use channel 1 and 4, respectively. This is a good choice since both not only detect each other but also detect AP2 and eNodeB 1 and choose different channels (from those selected by these detected nodes) for their transmissions.

In order to evaluate the usefulness of the proposed algorithm, the performance is measured in terms of the cumulative distribution function (CDF) of the normalized throughput. In Fig. 4 the CDF performance of the normalized throughput of the proposed scheme is compared with that of the optimum channel allocation algorithm. The optimum channel allocation assigns the channels to the transmission nodes in such way that the total network throughput maximizes and, it is found by exhaustive search over all possible channel combinations. Fig. 4 shows that the performance of the proposed Q-learning algorithm is close to the optimum one. Specifically, it is shown that under the proposed algorithm users have high probability to achieve the near-optimum normalized throughput with much less computational complexity.

V. Conclusion

In this paper, we have studied machine-learning-based channel selection in a coexistence scenario of LAA and Wi-Fi networks. We have proposed a reinforcement learning-based spectrum selection algorithm that allows the concurrent operating of Wi-Fi and LTE networks in an intelligent and efficient way. This scheme has taken into account the effects of both MAC and physical layers. We evaluated each channel's probability of selection, convergence property, and sum throughput. Simulation results demonstrate that this scheme converges fast and provides a competitive sum throughput reasonably close to that based on exhaustive search.

References

[1] “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017-2022 White Paper,” *Cisco Systems*, Feb. 2019.

[2] “Study on licensed-assisted access to unlicensed spectrum (Release 13),” *3GPP TR. 36.889 v13.0.0*, June 2015.

[3] S. Verma and S. Adhikari, “3GPP RAN1 status on LAA and NR Unlicensed,” *IEEE 802.11-18/0542r0*, Mar. 2018.

[4] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, Deep reinforcement learning for dynamic multichannel access in wireless networks, IEEE Trans. Cogn. Commun. Netw., vol. 4, no. 2, pp. 257265, Jun. 2018.

[5] O. Naparstek and K. Cohen, Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks, in Proc. GLOBECOM, pp. 17, Dec. 2017.

[6] M. A. Aref, S. K. Jayaweera, and S. Machuzak, Multi-agent reinforcement learning based cognitive anti-jamming, in Proc.IEEE Wireless Commun. Netw. Conf. (WCNC), pp. 16, Mar. 2017.

[7] X. He, H. Dai, and P. Ning, Faster learning and adaptation in security games by exploiting information asymmetry, IEEE Trans. Signal Process., vol. 64, no. 13, pp. 34293443, Jul. 2016

[8] V. Maglogiannis, D. Naudts, A. Shahid, and I. Moerman, A Q-learning scheme for fair coexistence between LTE and Wi-Fi in unlicensed spectrum, IEEE Access, vol. 6, pp. 2727827293, 2018.

[9] U. Challita, L. Dong, and W. Saad, Deep learning for proactive resource allocation in LTE-U networks, in Proc. 23rd Eur. Wireless Conf. Eur. Wireless, pp. 16, May 2017.

[10] Z. Liu, and I. Elhanany, “RL-MAC: a reinforcement learning based MAC protocol for wireless sensor networks,” *Int. J. of Sens. Netw. (IJSNET)*, Vol. 1, No. 3/4, 2007.

[11] O. Sallent, J. Prez-Romero, R. Ferrs, and R. Agust, “Learning-based Coexistence for LTE Operation in Unlicensed Bands,” *IEEE Int. Conf. on Commun. Wkshp*, June 2015.

[12] N. Rupasinghe, and I. Guvenc, “Reinforcement Learning for Licensed-Assisted Access of LTE in the Unlicensed Spectrum”, *IEEE Wireless Commun. and Netw. Conf.*, Track 3: Mobile and Wireless Netw., Mar. 2015.

[13] U. Challita, L. Dong, and W. Saad,“Deep Learning for Proactive Resource Allocation in LTE-U Networks”, *23th European Wireless Conf.*, Aug. 2017.

[14] “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures, (Release 15),” *3GPP TS 36.213 V15.6.0*, June 2019.

[15] IEEE Standard for Information technology, “Part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications,” *IEEE Std 802.11-2016 (Revision of IEEE Std 802.11-2012)*, pp. 1-23534, 2016.

[16] A. Babaei, J. Andreoli-Fang, Y. Pang, and B. Hamzeh, “On the impact of LTE-U on Wi-Fi performance,” *Int. J. of Wireless Inf. Netw.*, vol. 22, issue 4, pp. 336-344, Dec. 2015.

[17] E. Perahia and R. Stacey, “Next Generation Wireless LANs: 802.11n and 802.11ac,” Cambridge, U.K.: Cambridge Univ. Press, June 2013.

[18] G. Bianchi, “Performance Analysis of the IEEE 802.11 distributed coordination function,” *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535-547, Mar. 2000.

[19] Y. Ma, D. G. Kuester, J. Coder, and W. F. Young, “Slot-Jamming Effect and Mitigation Between LTE-LAA and WLAN Systems With Heterogeneous Slot Durations”, *IEEE Trans. on Commun.*, vol. 67, no. 6, pp. 4407-4422, June 2019.

[20] S. Mosleh, Y. Ma, J. B. Coder, E. Perrins, and L. Liu, “Enhancing LAA Co-existence Using MIMO Under Imperfect Sensing”, accepted for publication in the *IEEE GC Wkshps*, 2019.

[21] R.S. Sutton, A. G. Barto, “Reinforcement Learning: An Introduction,” MIT Press, 1998.

[22] S. Geman, D. Geman, “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. PAMI-6, No. 6, pp. 721-74, Nov. 1984.

[23] “Study on NR-based access to Unlicensed Spectrum; (Release 16),” *3GPP TR 38.889 V16.0.0*, Dec. 2018.

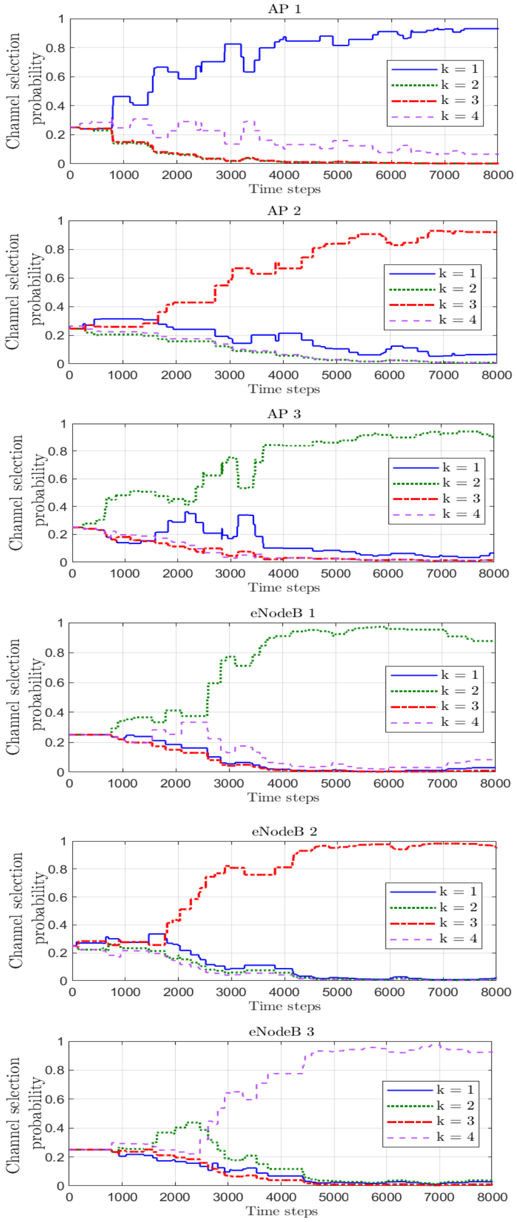


Fig. 3: Channel selection probability of each transmission node versus the time steps.

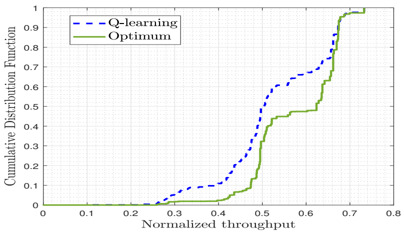


Fig. 4: CDF of achieved normalized throughput.

Mosleh, Susanna; Ma, Yao; Rezac, Jake; Coder, Jason. ”Dynamic Spectrum Access with Reinforcement Learning for Unlicensed Access in 5G and Beyond.” Paper presented at 2020 IEEE 91st Vehicular Technology Conference: VTC2020-Spring, Antwerp, BE. May 25, 2020 - May 28, 2020.

Influence of Angle Orientation on Firebrand Production from the Combustion of Materials Exposed to Applied Wind Fields

MANZELLO, Samuel L.[1], SUZUKI, Sayaka.[2]

[1] *National Instiute of Standards and Technology (NIST), USA*
[2] *National Research Institute of Fire and Disaster (NRIFD), Japan*

Abstract: A shared feature in the rapid spread of large outdoor fires are the production or generation of new, far smaller combustible fragments from the original fire source referred to as firebrands. A simplified experimental protocol has been developed that allows for the study of firebrand generation processes from various materials exposed to an applied wind field. The influence of angle of orientation on the firebrand production process is presented.

Keywords: Firebrands; Large Outdoor Fires; Ignition

1. Introduction

Across the globe, there exist many recent examples of large outdoor fires. Perhaps the most common are wildland fires that approach urban areas. These are often called Wildland-Urban Interface (WUI) fires. Two recent examples are the 2019 WUI fires that occurred in South Korea and those in 2018 in Northern California in the United States. In countries that are less developed, there have been large fires that have occurred in informal settlements. Some recent examples are those in the Philippines as well as South Africa, both in 2017. There have also been large urban fires in Japan, a country with no large wildland fire problem or informal settlement situation. Such urban fires have been recorded for centuries in Japan's history [**1**].

As communities become involved in large outdoor fires, firebrands are generated from the combustion of various structural fuel types. These include common building materials such as plywood, oriented strand board (OSB), and wood supporting members. At first glance, the liberation of firebrands from these combustion processes appears straight forward but is in fact quite complex. As a result, there is a need develop simplified experimental protocols to be able to investigate these combustion processes in a cost-effective manner. Such experiments have the possibility to aid in the development of computational models to understand these complex firebrand liberation processes from structural fuels. At the same time, such experiments may also lead to new insights into the physics of firebrand generation from various materials to perhaps suggest new methods of material fabrication to lessen firebrand liberation processes from these combustion reactions [**1**].

A simplified experimental protocol has been developed that allows for the study of firebrand generation processes from various materials exposed to an applied wind field. Preliminary results of this methodology was presented at this symposium last year, where it was shown useful insights into complex firebrand production processes was possible from such a simplified method [**2**]. To further understand firebrand liberation processes, mock-ups of actual structural assemblies were ignited and the angle of orientation of the burning assembly with respect to the wind was is studied.

2. Experimental Description

All experiments used mock-ups of structural assemblies and were performed in a wind facility at the National Research Institute of Fire and Disaster (NRIFD). NRIFD's wind facility has a 4 m diameter fan. The flow field was measured to be within ± 10 % over a cross-section of 2.0 m by 2.0 m. Experiments were performed within this cross-section.

Mock-ups used in this study were roofing assemblies, constructed with OSB and wood studs, with the dimensions of 610 mm (W) x 610 mm (H) (**Fig. 1**). Specific roofing treatments, such as roof tiles, were not applied but they are the obvious next step. These assemblies were exposed to applied wind fields and the angle of the assembly with respect to the wind was varied in increments of 25°, 45°, and 65 °. This was adjusted using a series of custom mounting frames fabricated for this study.

Experiments were performed in the following manner. Mock-up assemblies were ignited by a T-shaped burner with heat release rate (HRR) of 32 kW ± 10 % for 10 min. The reason to ignite under no wind was to provide consistent flame contact area for the assembly ignition for all experimental cases, independent of the wind speed. These afforded repeatable ignition conditions.

The total ignition time of 10 min was very important and carefully selected during ignition under no wind. If an ignition time less than 10 min was applied, once the wind field was added, the roofing assembly would self-extinguish, and little or no firebrand collection was possible due to lack of combustion. If an ignition time longer than 10 min was applied, the roofing assembly was observed to be consumed a great deal before the application of wind, thus rendering it unstable for firebrand collection (i.e. large holes were formed due to long ignition combustion process compromising the structural integrity).

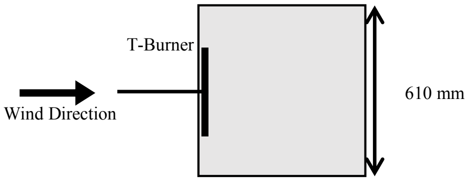


Figure 1 Schematic of roofing assemblies used for the experiments. Top view is shown (top view). OSB was used as base sheathing.

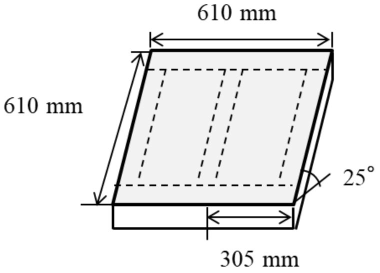


Figure 2 Schematic of mock-up assemblies used for the experiments (front view). The angle was varied from 25° to 65°. For illustration purposes, an angle of 25° is shown.

After the burner was turned off, a desired wind speed (6 m/s or 8 m/s) was applied. Experiments were stopped when the combustion of the assemblies was completed or the assemblies were not able to support themselves anymore.

A series of pans filled with water, placed downwind from the assemblies, were used to collect firebrands generated from the mock-ups. Water was needed to quench combustion otherwise only ash remained when experiments were finished. After the experiments, pans were collected, and firebrands were filtered and dried at 104 °C for 24 h. Dried firebrands were measured with a scale and pictures were taken for image analysis.

Manzello, Samuel; Suzuki, Sayaka. "Influence of Angle Orientation on Firebrand Production from the Combustion of Materials Exposed to Applied Wind Fields." Paper presented at 57 Japanese Combustion Symposium, Hokkaido, JP. November 20, 2019 - November 22, 2019.

3. Results and Discussion

Experiments were conducted for wind speeds of 6 m/s and 8 m/s. Commercially available image analysis software was used to determine the projected area of a firebrand by converting the pixel area using an appropriate scale factor. The projected areas with the maximum dimension of three dimensions were measured. The easiest method to visualize this is to imagine a sheet paper, as firebrands are often of very thin thickness, so the largest projected area is of interest. Images of specific shapes that have known areas (e.g. circles) were used to determine the ability of the image analysis method to calculate the projected area [2]. The standard uncertainty in determining the projected area was ± 10 %. Repeat measurements of known calibration masses were measured by the balance which was used for the firebrand mass analysis. The standard uncertainty in the firebrand mass was approximately ± 1 %.

Figs. 3-4 display a comparison of the collected firebrand size and mass distribution collected from these experiments. To more clearly show the results, linear curve fits are applied to the data. As may be seen, the angle of orientation had very little influence on the size and mass of liberated firebrands for an applied wind speeds of 6 m/s. As the wind speed was increased, for a given projected area, the mass of firebrands generated was quite sensitive to the angle of orientation.

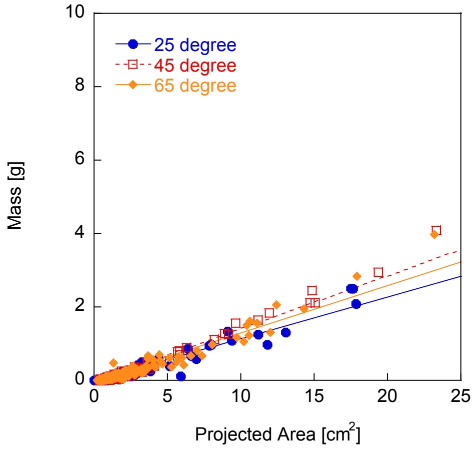


Figure 3 The mass and the projected Area of firebrands collected under 6 m/s. Data is shown for structural assemblies oriented at three different angles with respect to the applied wind speeds.

It is known that the wind force plays an important role in liberating firebrands from structural fuels. In its most simplistic representation, the wind force is proportional to the square the applied wind speed. As the projected area of liberated firebrands has a linear relationship, the values of slopes presented in **Figs. 3-4** are summarized in **Fig. 5**. The mass of a given firebrand may be expressed as follows:

$$m_F = \rho_F d A_{proj} \qquad (1)$$

where m_F is the firebrand mass, ρ_F is the firebrand density, d is the firebrand thickness, and A_{proj} is the projected area. As all the firebrands are liberated from the same materials in these experiments, it is reasonable to assume the firebrand densities are similar. Therefore, the largest slope indicates firebrands with the largest thickness, suggesting that the thickest firebrands were produced at angles of 45° for applied wind speeds of 8 m/s.

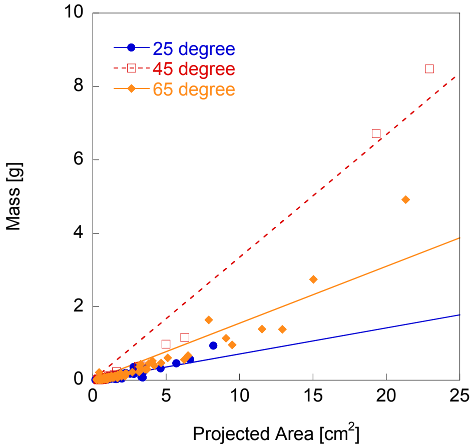


Figure 4 The Mass and the projected Area of firebrands collected under 8 m/s. Data is shown for structural assemblies oriented at three different angles with respect to the applied wind speeds.

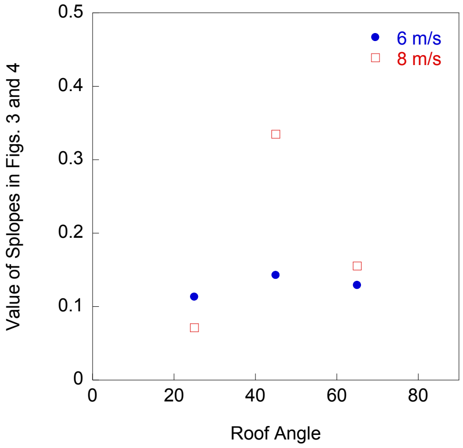


Figure 5 Effect of wind speeds on characteristics of firebrands generated from structural assemblies (mock-up roofing).

4. Conclusions

A simplified experimental protocol has been developed that allows for the study of firebrand generation processes from various materials exposed to an applied wind field. At wind speeds of 8 m/s, the angle of orientation greatly influenced the characteristics of the liberated firebrands.

5. Acknowledgments

Mr. Marco Fernandez of NIST's Engineering Laboratory is appreciated for sending experimental supplies that made this work possible.

References

1. Manzello, S.L., Suzuki, S., Gollner, M.J, and Fernandez-Pello, A.C., *Progress in Energy and Combustion Science*, in press, 2019.
2. Manzello, S.L., and Suzuki, S., Proceedings of the 56th Japanese Combustion Symposium, Osaka, Japan, 2018.

Manzello, Samuel; Suzuki, Sayaka. "Influence of Angle Orientation on Firebrand Production from the Combustion of Materials Exposed to Applied Wind Fields." Paper presented at 57 Japanese Combustion Symposium, Hokkaido, JP. November 20, 2019 - November 22, 2019.

Investigating coupled effect of radiative heat flux and firebrand showers on ignition of fuel beds

SUZUKI, S. [1], MANZELLO, Samuel L.[2]

[1] *National Research Institute of Fire and Dissaster (NRIFD), Japan*
[2] *National Institute of Standards and Technology (NIST), USA*

Abstract: Fire spread occurs via radiation, flame contact, and firebrands. While firebrand showers are known to be a cause of spot fires which ignite fuels far from the main fire front, in the case of short distance spot fires, radiation from the main fire may play a role for firebrand induced ignition processes. Many past investigations have focused on singular effects on fire spread, and little is known about coupled effects. The coupled effect of radiative heat flux and firebrand showers on ignition processes of fuel beds is studied by using a newly developed experimental protocol. Experiments were performed under the an applied wind field, as the wind is a key parameter in large outdoor fire spread processes

Keywords: Firebrands; Radiation, Large Outdoor Fires; Ignition

1. Introduction

Large outdoor fires pose problems for societies across the world. Perhaps the most often in the news are wildland fires that approach urban areas. These are more simply referred to as Wildland-Urban Interface (WUI) fires. In Asia and North America, some recent examples are the 2019 WUI fires that occurred in South Korea and those in 2018 in Northern California in the United States. In Africa, in less developed countries, there have been large fires that have occurred in informal settlements. For centuries there have also been large urban fires in Japan, a country with no large wildland fire problem or informal settlement situation [**1**].

In all of these large outdoor fires, firebrands are produced and lead to enhanced fire spread processes. Often referred to as spotting processes, firebrands are liberated from the combustion of various fuel types and then induce ignition of fuel sources away from the initial fire source. A potentially important aspect of the physics of ignition induced by firebrand are the coupled influence of firebrand showers and radiant heat.

Here, the coupled effect of radiation and firebrand showers on ignition processes of fuel beds is studied by using a newly developed experimental protocol. Experiments were performed under an applied wind field, as the wind is a key parameter in large outdoor fire spread processes

2. Experimental Description

All experiments were conducted at the National Research Institute of Fire and Disaster (NRIFD). To generate firebrand showers, the reduced-scale continuous-feed firebrand generator was used and installed inside NRIFD's wind facility. The device consisted of two parts; the main body and continuous feeding component.

A conveyer was used to feed wood pieces continuously into the device. The conveyer belt was operated at 1.0 cm/s, and wood pieces were put on the conveyer belt at 12.5 cm intervals. Douglas-fir wood pieces machined with dimensions of 7.9 mm (H) by 7.9 mm (W) by 12.7 mm (L) were used to produce firebrands. These same size wood pieces were used in past studies and have been shown to be within projected area/mass of firebrands measured from full-scale burning trees as projected areas obtained from actual WUI fires [**2**]. The wood feed rate used here was 80 g/min, which is near the upper limit for this reduced-scale firebrand generator [**2**].

As the base of the fan used to generate the wind in the NRIFD facility is located 1. 6 m from the floor, the conveyer was placed under a custom stage designed for experiments when using NRIFD's wind facility. The wind field exits from a 4.0 m diameter fan, and it is possible to generate wind speeds up to 10 m/s. The flow field was measured to be within ± 10 % over a cross-section of 2.0 m by 2.0 m.

When the blower was set to provide an average velocity below 5.0 m/s measured at the exit of the firebrand generator when no wood pieces were loaded, insufficient air was supplied for combustion and this resulted in a great deal of smoke being generated in addition to firebrands. Above 5.0 m/s, smoke production was mitigated but then many firebrands produced were in a state of flaming combustion as opposed to glowing combustion. In these experiments, glowing firebrands were desired.

The fuel beds used for ignition were 300 mm by 300 mm in size and consisted of the same Douglas-fir wood pieces used to generate firebrands (**Fig. 1**). These were installed inside a mock-up corner assembly lined with calcium silicate board, since the ignition of the corner assembly itself was not the goal here; only ignition induced in the wood pieces was of interest.

To provide uniform radiant heat flux to fuel beds, an electrically operated quartz radiant panel was used. Dimensions of the radiant panel were also 300 mm by 300 mm (**Fig. 1**). It was mounted at a height of 440 mm from the fuel bed surface. A custom calibration rig was designed and fabricated to quantify the radiant heat flux that the radiant panel provided at the fuel bed surface. Pre-heating time was determined as the duration from the time when the radiant panel was turned on to the time to start the firebrand generator. Selected pre-heating time was 0 min, 10 min, and 20 min. Baseline experiments were also performed to examine the ignitions by firebrands without applied radiant heat flux. The two wind speeds were selected, namely 6 m/s and 8 m/s.

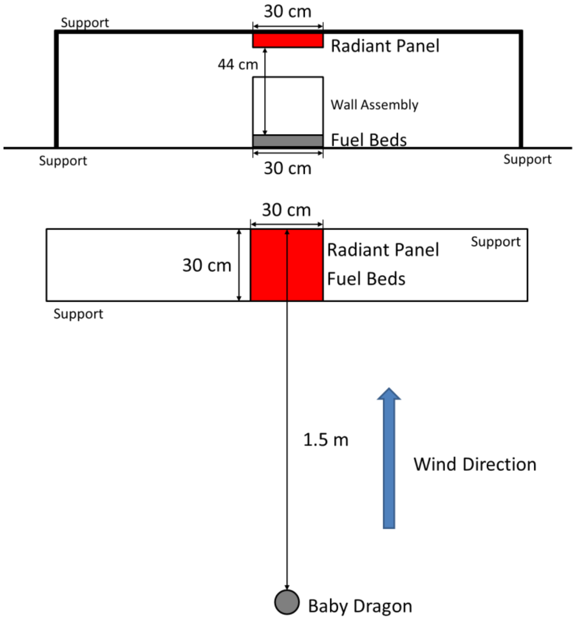


Figure 1 Schematic of the experimental settings. Front view (top) and Top view (bottom) are shown. Firebrand showers were produced using the reduced-scale firebrand generator installed in NIRFD's wind facility. Fuel beds consisted of Douglas-fir pieces of uniform size.

Manzello, Samuel; Suzuki, Sayaka. "Investigating coupled effect of radiative heat flux and firebrand showers on ignition of fuel beds." Paper presented at 57 Japanese Combustion Symposium, Sapporo (Hokkaido), JP. November 20, 2019 - November 22, 2019.

3. Results and Discussion

Fig. 2 displays a picture of a typical experiment. In this picture, the applied wind speed was 8 m/s and radiant panel was switched on at the same time as firebrand generator, that is there was no pre-heating time.

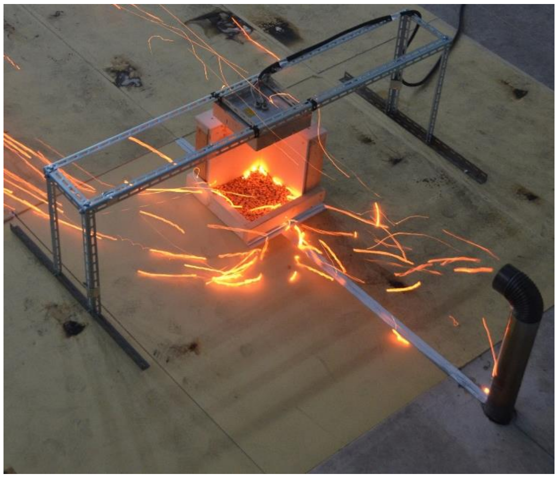


Figure 2 A typical experiment conducted at 8 m/s. Multiple glowing firebrands are seen depositing in the fuel bed.

Fig. 3 displays the time to smoldering ignition in the fuel beds a function of pre-heating time and wind speed. Here, sustained smoldering ignition (SI) was defined as the start of intense smoke generation in the fuel be due to the coupled effects of accumulated firebrands and radiant heat flux. To aid in understanding, in **Fig 3**, no radiant panel conditions imply that the radian panel was not switched on. For other cases, the applied radiant heat flux to fuel beds was 10 kW/m^2.

An interesting result from these findings is that for wind speed of 6 m/s, the pre-heating due to the applied radiant heat flux to the fuel bed influenced the time to SI significantly. Yet, as the wind speed was increased to 8 m/s, the applied radiant heat flux had very little effect.

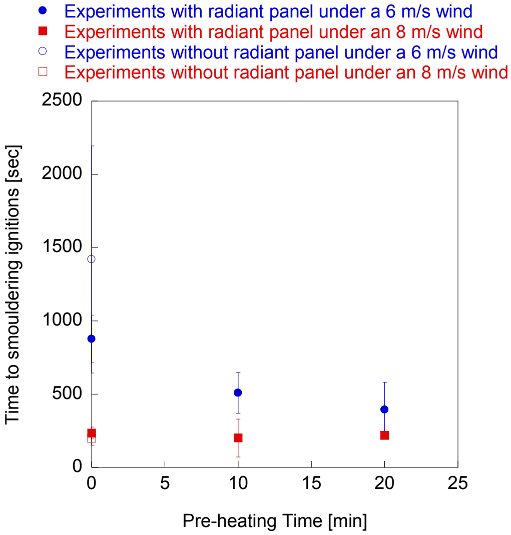


Figure 3 Coupled effects of radiation and firebrands on ignition under different wind speeds and pre-heating time.

The ignition time for thermally thick materials is known to be proportional to the room-temperature density, ρ, of the material and inversely proportional to the square of the net heat flux to the fuel bed [**3**]. As a result, at 6 m/s, the applied radiant heat flux acts to reduce the time to SI ignition greatly with increased pre-heating time. As wind speed is increased, convection cooling to the fuel beds is increased but the time to ignition was not influenced by the additional assistance from the applied radiant heat flux. It is believed that the increased wind speed rather changes the combustion dynamics of the firebrands and therefore alters their temperatures, resulting in larger applied heat fluxes from the firebrands themselves.

Manzello *et al.* [**3**] quantitatively showed that glowing firebrand surface temperature increased as the wind speed was increased. In that work, the surface temperature of charred area on the glowing firebrand was simultaneously measured using both the thermocouple and the infrared camera [**3**]. The glowing, charred, area close to the thermocouple bead was imaged using the infrared camera and the emissivity was then adjusted until the temperature readings on the infrared camera converged to those registered by the thermocouple measurements. Temperatures measured by the infrared camera approached those measured by the thermocouple as the emissivity set on the camera was reduced from 0.7 to 0.6. The firebrand temperatures were quite sensitive to airflow (**Fig. 4**). Therefore, larger firebrand temperatures were expected as the wind speed was increased from 6 m/s to 8 m/s.

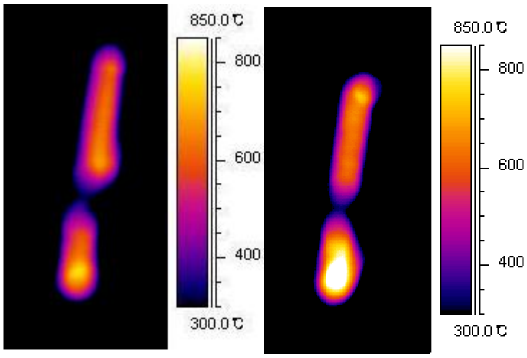


Figure 4 IR images of the glowing firebrand exposed to two different air flows (1.3 m/s, left and 2.4 m/s, right) [**3**].

4. Conclusions

The coupled effect of radiative heat flux and firebrand showers on ignition processes of fuel beds was studied by using a newly developed experimental protocol. An interesting result from this preliminary study is that for wind speed of 6 m/s, the pre-heating time due to the applied radiant heat flux to the fuel bed influenced the time to SI significantly. Yet, as the wind speed was increased 8 m/s, the applied radiant heat flux had very little effect. It was postulated that differences in firebrand temperatures may explain these differences. More work is required to quantify firebrand temperatures at higher wind speeds.

5. Acknowledgments

This research is supported by MEXT/JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 18K04675.

References

1. Manzello, S.L., Suzuki, S., Gollner, M.J, and Fernandez-Pello, A.C., *Progress in Energy and Combustion Science*, in press, 2019.
2. Suzuki, S., and Manzello, S.L., Proceedings of the 53th Japanese Combustion Symposium, Tsukuba, Japan, 2015.
3. Manzello, S.L., Park, S.H., and Cleary, T.G., *Fire Saf. J.* 44: 894-900, 2009.

Manzello, Samuel; Suzuki, Sayaka. "Investigating coupled effect of radiative heat flux and firebrand showers on ignition of fuel beds." Paper presented at 57 Japanese Combustion Symposium, Sapporo (Hokkaido), JP. November 20, 2019 - November 22, 2019.

Forensic Analysis of Advanced Persistent Threat Attacks in Cloud Environments

Changwei Liu[1], Anoop Singhal[2], Duminda Wijesekera[1,2]

[1]Department of Computer Science, George Mason University, Fairfax VA 22030 USA
[2]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg MD 20899 USA
[1]{cliu6,dwijesek}@gmu.edu ,[2]{anoop.singhal}@nist.gov

Abstract: Due to the increasing cyber-activities and the use of diverse devices offered on cloud environments, cloud forensic investigations must deal with data in diverse formats and large quantities from different devices. The process of forensic investigation in a cloud environment involves filtering away noisy data and using expert knowledge to make up the missing attack steps from advanced persistent threats (APT) attacks that have a long time span. Under such circumstance, obtaining timely and convincing forensic results is a challenge.

We show how MITRE's ATT&CK framework and Lockheed Martin's cyber kill chain can be used to identify forensically valuable data, aggregate and correlate them to construct attack steps. ATT&CK is a globally accessible knowledge base of adversary tactics and techniques that is based on real-world observations. APT attacks on cloud systems consist of a successful reconnaissance, command and control communication, privilege escalation, lateral movement through the network, exfiltration of confidential information that are also the key phases of a cyber kill chain. In this paper we investigate using cyber kill chains to organize the evidence and construct the attack steps. By using an experimental network, we show how our methodology can be used to identify evidence, aggregate them and feed them to a Prolog-based tool to re-construct attack steps for the purpose of performing forensic analysis.

Keywords: Cloud attacks, forensic analysis, ATT&CK, cyber kill chain, attack steps

1. Introduction

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda. "Forensic Analysis of Advanced Persistent Threat Attacks in Cloud Environments." Paper presented at Sixteenth IFIP 11.9 International Conference on Digital Forensics, New Delhi, IN. January 06, 2020 - January 08, 2020.

Cyber forensics is the application of science to the identification, collection, examination, and analysis, of data while preserving the integrity of the information and maintaining a strict chain of custody [21]. Due to a continued increase in cyber-activities and the diversity of devices that use services provided in cloud environments, the scope of post attack forensic investigations, especially the ones involving attacks on cloud environments have expanded in two dimensions: (1) Expansion of the attack surfaces created by cloud devices that may not have undergone rigorous security check; (2) The need to analyze data that comes in diverse formats and quantities from new attackable interfaces. Of special concern is that instances of servers running on virtual machines in the cloud are monitored by hypervisors that lack warnings, procedures and tools for forensic investigations. Because existing forensic techniques have not been designed for cloud environments, it is challenging to use existing tools to perform forensic analysis in a cloud environment, especially for those advanced persistent threat attacks (APTs) that could last for so long(such as a year) that the timestamps of evidence are not the indicators of the same attack. The process of investigating such cloud attacks involves filtering away noisy data and using expert knowledge or experience to speculate the attacker strategy. This situation creates increasing obstacles in their abilities to obtaining pertinent forensic results.

Researchers have proposed methodologies to collect evidence and correlate them for forensic analysis of cloud attacks, which include developing tools to collect data from hypervisors or virtual machines (VM)s [5,6], leveraging graphical framework to reconstruct the attack scenarios [2, 7] etc. However, cited research is based on strong assumptions that the forensic data was manually aggregated and preprocessed to evidence representing pre-attack conditions and post-conditions, and uses forensic investigators' expertise to build an attack step when the corresponding evidence is incomplete or compromised.

Recently, ATT&CK-like frameworks have been used in understanding and fighting cyber-attacks. Lockheed Martin's intrusion kill chain (also called *cyber kill chain*) describes seven phases including reconnaissance, weaponization, delivery, exploitation, installation, command and control, actions on objectives. Because most APT attacks consist of a successful reconnaissance, command and control communication, privilege escalation, lateral movement through the network, exfiltration of confidential information etc., the cyber kill chain has been used to analyze security logs, develop attack detection and defense systems, and aggregate evidence for APT attack forensic analysis [1,3]. MITRE ATT&CK is a globally accessible knowledge base of adversary tactics and techniques based on real-world observations [11] to emulate cyber-attacks. It has been used in recent years to create a taxonomy of possible attacks in the enterprise IT environment to allow defenders to understand what attacks are being used in the wild and provide methods to detect attacks including certain APTs. Inspired by these works, we propose to leverage MITRE's Adversarial Tactics and Common Knowledge (ATT&CK) and Lockheed Martin's cyber kill chain to identify the evidence of cloud APT attacks, aggregate it and correlate it to construct the attack steps, as the enhancement to our previous work [2,7] of performing forensics in a cloud environment, which had to depend on the investigator's experience and knowledge to identify evidence and building the attack step when the corresponding evidence is incomplete or compromised. Though existing research uses the two tools to detect cyber-attacks or aggregate/correlate evidence [1,3, 19, 20], to the best of our knowledge, there is no published research work that combines both for attack evidence identification and correlation. That is the main contribution in this paper. We use sample attacks to show how both frameworks can be used to identify forensic data in a cloud environment, covert it to pre-attack and after-attack evidence, and feed it to a logic-based forensic tool to construct attack steps.

The rest of the paper is organized as follows. Section 2 describes the background knowledge and related work. Section 3 shows the experimental attacks and the collected evidence. Section 4 describes how we use the two frameworks to identify, correlate evidence, and implement an existing Prolog based forensic tool to construct attack steps. Finally, we conclude the paper in Section 5.

2. Background Knowledge and Related Work

We describe Lockheed Martin's cyber-kill chain model, MITRE ATT&CK framework and related work in this section.

2.1 The Cyber-Kill Chain Model

As shown as the bottom part in Figure 1, there are seven attacking phases in Lockheed Martin's kill-chain model, representing the steps composed of a successful network attack. In "reconnaissance" phase, the adversaries identify the targets by researching what targets can meet their attacking objectives, and correspondingly collect information that could enable them to launch attacks; in "weaponization" phase, the adversaries prepare their operations by coupling malware and exploits into a deliverable payload, selecting backdoors/implants and appropriate command and control infrastructures for operations; in "delivery" phase, the adversaries convey the malware to the target to launch the attack operations; in "exploitation" phase, the adversaries trigger exploits to gain access to the targeted victim; in "installation" phase, the adversaries install a persistent backdoor or an implant in the victim to maintain the access for an extended period; in "command and control(C2)" period, the adversaries remotely control the backdoor or implant to open a command channel so that the adversaries could control the victim; in the last phase, "actions on objectives", the adversaries achieve their attacking objectives that include collecting user credentials, escalating their privileges, destroying the system, overwriting or corrupting data or modifying data [8]. According to Milajerdi at el.[1], most APT attacks are accomplished through steps conforming to the cyber kill chain, and aim to obtain and exfiltrate highly confidential information.

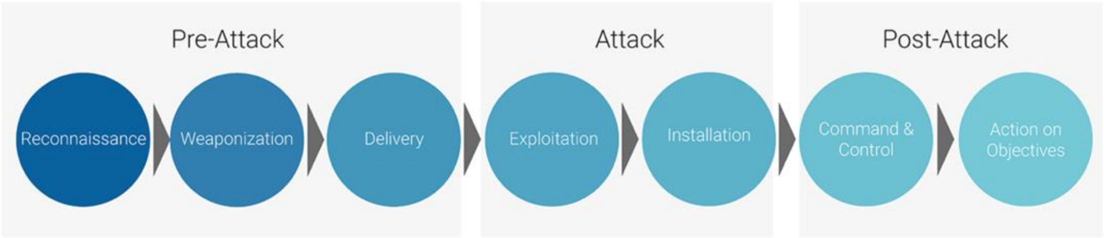


Figure 1. The cyber intrusion kill chain from Lockheed Martin

Cyber Reboot (an American company) reexamined the seven phases and argued that there are three fundamental phases to most network attacks. They are pre-attack, attack and post-attack phases as shown at the top of Figure 1. In the pre-attack phase, the attacker is tasked with attacking objectives and starts to perform reconnaissance of the target; in the attack phase, the attack is executed, enabling the attacker to break through the target's defense and set up communication between the attacked target and the attacker; in the post-attack phase, a further exploitation or access of the targeted victim occurs, which allows the attacker to escalate privilege, destroy the victim system, steal confidential information and etc.

2.2 The Adversarial Tactics and Common Knowledge (ATT&CK) Framework

The ATT&CK framework proposed by MITRE [19] is a behavioral model based on real-world observations. As shown in Figure 2, it uses the cyber kill chain to emulate cyber-attacks by emulating the tactics and techniques the attackers could use to achieve their goals. Figure 3 is a snippet of MITRE ATT&CK matrix of enterprise networks, from which we can see that, unlike other threat models that were built by analyzing available threat/vulnerability reports, ATT&CK describes behaviors commonly employed by real adversaries. All attacking techniques are real-world examples of malware or from threats used by red teams. Besides, this framework has public descriptions of attack techniques and how they are leveraged and why the network defenders should pay attention. Therefore, it is useful for network defenders and forensic investigators to decide what they should monitor and how to specifically investigate to reconstruct the attack steps and mitigate the attack risks.

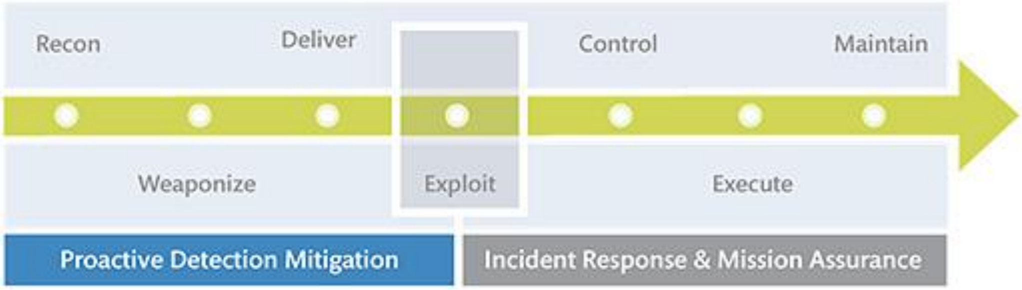
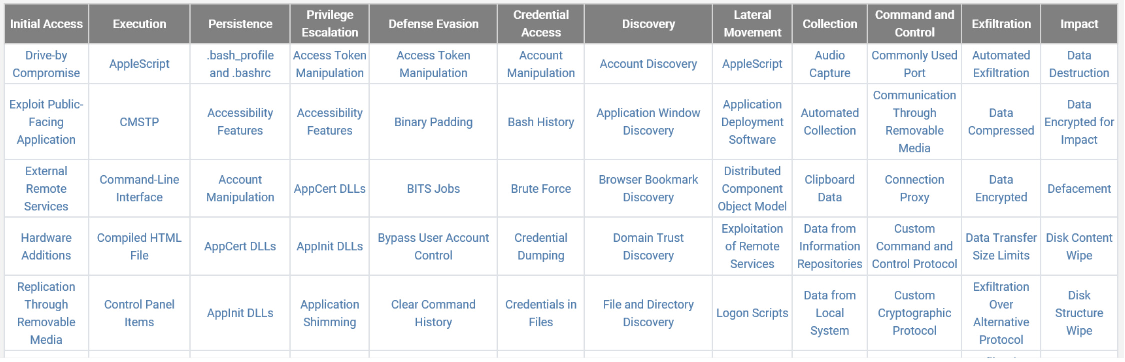

Recon
Deliver
Control
Maintain
Weaponize
Exploit
Execute
Proactive Detection Mitigation
Incident Response & Mission Assurance

Figure 2. MITRE ATT&CK framework



Initial Access	Execution	Persistence	Privilege Escalation	Defense Evasion	Credential Access	Discovery	Lateral Movement	Collection	Command and Control	Exfiltration	Impact
Drive-by Compromise	AppleScript	.bash_profile and .bashrc	Access Token Manipulation	Access Token Manipulation	Account Manipulation	Account Discovery	AppleScript	Audio Capture	Commonly Used Port	Automated Exfiltration	Data Destruction
Exploit Public-Facing Application	CMSTP	Accessibility Features	Accessibility Features	Binary Padding	Bash History	Application Window Discovery	Application Deployment Software	Automated Collection	Communication Through Removable Media	Data Compressed	Data Encrypted for Impact
External Remote Services	Command-Line Interface	Account Manipulation	AppCert DLLs	BITS Jobs	Brute Force	Browser Bookmark Discovery	Distributed Component Object Model	Clipboard Data	Connection Proxy	Data Encrypted	Defacement
Hardware Additions	Compiled HTML File	AppCert DLLs	AppInit DLLs	Bypass User Account Control	Credential Dumping	Domain Trust Discovery	Exploitation of Remote Services	Data from Information Repositories	Custom Command and Control Protocol	Data Transfer Size Limits	Disk Content Wipe
Replication Through Removable Media	Control Panel Items	AppInit DLLs	Application Shimming	Clear Command History	Credentials in Files	File and Directory Discovery	Logon Scripts	Data from Local System	Custom Cryptographic Protocol	Exfiltration Over Alternative Protocol	Disk Structure Wipe

Figure 3. A Snippet of MITRE matrix for enterprise networks

2.3 Related Work

Techniques, including remote data acquisition, management plane acquisition, live forensics and snapshot analysis, have been proposed to collect evidence from cloud environments [12]. Dykstra and Sherman retrieved volatile and non-volatile data from the Amazon EC2 cloud active user instance platform using traditional forensic tools such as EnCase and FTK [13]. In order to validate the integrity of the collected data, they subsequently developed the FROST toolkit that can be integrated within OpenStack to collect logs from an operating system that runs the virtual machines [14]. While this technique assumes that the cloud provider is trustworthy, Zawoad et al. resolved this issue by designing a forensics-enabled cloud [15]. Hay and Nance [16] have conducted live digital forensic analyses on clouds with virtual introspection, a process that enables the hypervisor or any other virtual machine to observe the state of a chosen virtual machine. Dolan-Gavitt et al. bridged the semantic gap between the high-level state information and low-level sources such as physical memory and CPU registers and developed a suite of virtual introspection tools for Xen and KVM [17]. Several hypervisors, including Xen, VMWare, ESX and Hyper-V, support snapshot features that can be used to obtain information about the running state of the virtual machine.

Meanwhile, in order to reduce the time and effort involved in forensic investigations, researchers have proposed automating evidence correlation and attack reconstruction by leveraging rule-based tools and business process diagrams [2]. However, this work depends on forensic experts when the evidence is missing, disjointed or compromised. To execute security investigators in a methodical manner and help to detect real-time APT attacks, intrusion kill-chain has been modified to facilitate data aggregation within a relational database [1, 3].

3. Experimental Attacks in a Cloud Environment

In this section, we show an experimental network attacked by both conventional and cloud cyber-attacks to show how the two frameworks of ATT&CK and cyber-kill chain are used to assist cloud forensic analysis.

3.1 Our Experimental Cloud and Sample Attacks

Based on the location of an vulnerability and the source of an attacker, the attacks in the cloud can be categorized to two groups [5, 7]: (1) an attacker from the Internet uses conventional network vulnerabilities to attack a virtual machine connected to the Internet; (2) an attacker from a virtual machine uses the vulnerabilities from the shared cloud management resources to launch attacks to other virtual machines on the same hypervisor. The attacks from both kinds of attacks range from Denial of Service, information leakage to privilege escalation, arbitrary code execution etc.

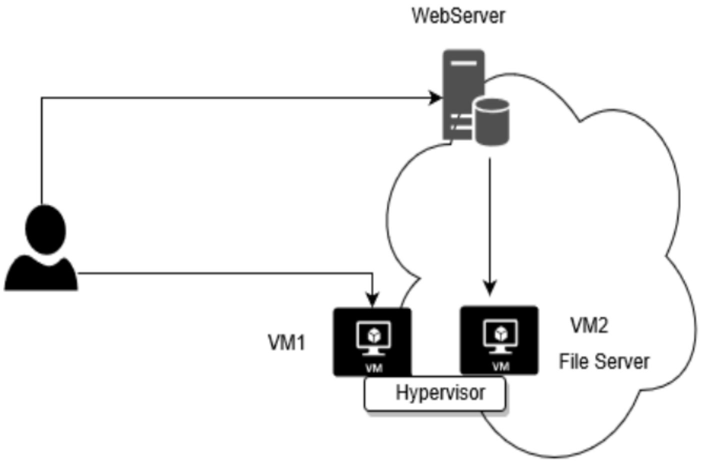


Figure 4. Our experimental cloud and sample attacks toward

Figure 4 illustrates our experimental set up in a lab environment. In this small cloud environment, VM1 and VM2 are two Linux (Ubuntu 14.04) virtual machines that are configured on the same hypervisor (Xen 4.6). We also configured a Windows machine as a webserver, in which a web application can use SQL queries to retrieve the database information stored on VM2--the file server that hosts databases and other files. Toward VM2, we launched two attacks. One is a conventional SQL injection attack exploited by using the vulnerability from the web application that does not sanitize the users' input. The other one is a VM escape attack that could be a kind of APT attacks. The VM escape attack takes advantages of the vulnerability CVE-2017-7228 from VM1, which allows VM1 to control Xen privileged domain, domain 0, and then VM2 so that it can perform local operations, such as deleting a file, in VM2.

3.2 Forensic Data Obtained by Using Forensic Tools

To obtain data for forensic analysis, during the attack processes, we logged the accesses toward the webserver, deployed Snort as an IDS to monitor network traffics to the webserver and file server, and installed a VM introspection tool, LibVMI, on Xen Dom0 to capture events and running processes in the guest VMs(VM1 and VM2). LibVMI is a C library that can be used to monitor the low-level details of a running VM of Xen by viewing its memory, trapping on hardware events, and accessing the vCPU registers. Below are the machine IP addresses and forensic data captured by using the methods/tools mentioned here.

Table 1. The IP address for each machine or VM in Figure 4

Machine/VM	IP Address
Attacker	129.174.124.122
Web Server	129.174.125.35
VM1	129.174.124.184
VM2(File server)	129.174.124.137

```
[**] SQL Injection Attempt --1=1 [**]
08/08-14:37:27.818279 c:1715 -> 129.174.124.35:8080
TCP TTL:128 TOS:0x0 ID:380 IpLen:20 DgmLen:48 DF
******S* Seq: 0xDEDBEABF  Ack: 0x0  Win: 0xFFFF  TcpLen: 28
TCP Options (4) => MSS: 1460 NOP NOP SackOK
… …
```

Figure 5. A sample Snort Alert

```
129.174.124.122 - - [08/Aug/2019:14:35:34 -0400] "GET /lab/Test HTTP/1.1" 200 368
129.174.124.122 - - [08/Aug/2019:14:35:39 -0400] "POST /lab/Test HTTP/1.1" 200 981
..
```

Figure 6. Sample access logging from the Web Server

```
130808 14:37:29         …
                    40 Query    SET NAMES latin1
                    40 Query    SET character_set_results = NULL
                    40 Query    SET autocommit=1
                    40 Query    SET GLOBAL general_log = 'ON'
                    40 Query    select * from profiles where name='Alice' AND password='alice' or
'1'='1'
                    40 Quit
```

Figure 7. Sample SQL database logging

```
…
[  630] agetty (struct addr:ffff880003c8e200)
[  669] systemd (struct addr:ffff880076060000)
[  674] (sd-pam) (struct addr:ffff880076104600)
[  677] bash (struct addr:ffff880003c8aa00)
[  703] sudo (struct addr:ffff880004341c00)
[  704] attack (struct addr:ffff880004343800)
```

(a) Running processes

```
test
intel_rapl
x86_pkg_temp_thermal
coretemp
….
```

(b) Injected Linux modules

```
Waiting for events...
PID 0 with CR3=77130000 executing on vcpu 1. Previous CR3=788d1000
Waiting for events...
PID 1246 with CR3=788d1000 executing on vcpu 1. Previous CR3=77130000
```

(c) CPU register values

Figure 8. The Processes, injected Linux modules, and CPU register values of VM2

Table 1 shows the corresponding IP addresses of each machine/VM. According to Table 1, we can see the Snort alert in Figure 5 shows the attacker from 129.174.124.122 attempted to launch a SQL injection attack by using the web application deployed in the webserver with IP address 129.174.125.35 at port number 8080. Figure 6 is the web access history on the webserver, which shows that the attacker machine accessed the web application right before the time when the Snort sent out the alert listed in Figure 5. Figure 7 is the SQL access logging that includes the SQL injection query (the line of “40 Queryselect * from profiles where name='Alice' AND password='alice' or '1'='1'”) that resulted in information leakage. Figure 8 are the results obtained by running LibVMI on the attacker VM. These results are forensic data without timestamps. They are composed of the running processes, injected Linux modules and CR3 register values representing the running processes (the process identifier (PID) is used to find the process name).

4. Leveraging the ATT&CK Matrix and Cyber Kill Chain for Forensic Investigation

In this section, we show how we used ATT&CK and the cyber kill chain to assist the investigation process.

4.1 Use ATT&CK to Identify the Forensic Data

MITRE ATT&CK includes a knowledge base of 11 tactics and hundreds of techniques an attacker would leverage when compromising an enterprise network. In this framework, while a tactic is a high-level description of a certain type of attack's behaviors, a technique provides a more detailed description of every specific type of behavior within a certain tactic class. The tactics from ATT&CK aren't followed in any linear order as Lockheed's cyber kill chain as shown in Figure 1, and an attacker could bounce between tactics in order to achieve his final goal.

We map forensic data to ATT&CK matrix in order to assist identifying evidence for cloud forensic investigation. As shown in Figure 3, the matrix model's various phases of an attack's lifecycle include "initial access", "execution", "persistence", "privilege escalation", "defense evasion", "credential access", "discovery", "lateral movement", "collection", "command and control", "exfiltration", and "impact". Each of these phases is consisted of different techniques as listed in the matrix table, and the general description is provided below.

- Initial Access consists of techniques that use various entry vectors to gain their initial foothold within a network, which may allow for the attacker's continued access to external remote services.
- Execution consists of techniques resulting in adversary-controlled code running on a local or remote system, which can achieve broader goals such as exploring a network or stealing data by paring with other technologies. Notice that this step does not leave any evidence.

- Persistence consists of techniques that the attacker uses to maintain their foothold on systems even if any interruptions from the system cut off their access.
- Privilege Escalation consists of techniques that an attacker uses to gain higher-level permissions on a system or network. The common approaches include taking advantage of system weaknesses, misconfigurations, and vulnerabilities.
- Defense Evasion consists of techniques such as uninstalling/disabling security software or obfuscating/encrypting data and scripts that an attacker uses to avoid detection throughout their compromise.
- Credential Access consists of techniques that an attack uses to steal credentials to gain access to systems, which provides the opportunity for the attacker to achieve his goals by creating more accounts.
- Discovery consists of techniques an attacker may use to gain knowledge about the system and internal network, which allows the attacker to explore what he can control and discover how the knowledge could benefit his post-compromise information-gathering objective.
- Lateral Movement consists of techniques that an attacker uses to enter and control remote systems on a network. An attacker might install his own remote access tools to accomplish lateral movement or use legitimate credentials with native network and operating system tools.
- Collection consists of techniques an attacker uses to gather information. Frequently, the next goal after collecting data is to steal the data.
- Command and Control consists of techniques that an attacker may use to communicate with systems under his control within a victim network.
- Exfiltration consists of techniques that an attacker may use to steal data from the victim network. The attacker often compresses or encrypt the data to avoid detection, and the channels used to get

data out of a victim network typically include the attacker's command and control channel or an alternate channel with limited size.

- Impact consists of techniques that an attacker uses to disrupt availability or compromise integrity by manipulating business and operational processes, which include destroying or tampering with data.

In our experimental network, we can identify the evidence of the SQL injection attack, because the Snort alert (Figure 5) and SQL query (Figure 7) of this attack clearly show it was a SQL injection attack. For the VM escape attack of using the vulnerability CVE-2017-7228, thought the attack could be observed (the file in VM2 has been deleted), it is hard to construct the attack from the data obtained by using LibVMI (Figure 8), because there was no obvious logging that can identify this attack. Clearly, this is a step that does not leave any evidence. Under this situation, using ATT&CK could narrow down the search scope, helping to find the evidence. According to ATT&CK, the initial access techniques include "drive-by compromise, exploit public facing application, external remote services, hardware additions, replication through removable media, spear phishing attachment, …, trusted relationship". In our experimental cloud, except for the facts that the database in VM2 can be queried by using the web application in the webserver and VM2 shares the same hypervisor (therefore, the hardware) with VM1, it does not have any other connected media, remote services or running applications. Thus, the initial accesses could be narrowed down to the tactics including "exploit public-facing application" from the webserver, and "hardware additions" from the hypervisor. Because the observed attack activities on VM2 include SQL injection alerts and a file deletion, according to ATT&CK, the execution techniques of the attacks could fall into two categories, "exploitation for client execution"(corresponding to application from the webserver) and "command-line interface"(corresponding to the hardware addition). And, the techniques for "privilege escalation" would be narrowed down to "exploitation for privilege escalation", since the attackers obviously escalated their privileges from the Internet or other VM remotely and other techniques including "access token manipulation", "accessibility features", … etc. are not suitable for the system configuration.

The SQL injection attack has obvious evidence as listed in Figure 5, 6 and 7. However, for the VM escape attack that resulted in the file deletion, the data provided in Figure 8 could only show all running process (including the normal Linux processes and the suspicious user process named as "attack") and injected modules (including normal Linux modules and the suspicious injected user module named as "test"), which does not show the process of exploiting the vulnerability that uses the shared hardware. Therefore, using the potential attack tactics obtained from ATT&CK, the forensic investigators should continue to investigate more forensic data related to exploitations that takes advantage of tactics including "hardware additions", "command-line interface" to launch attacks that allowed the attacker to escalate privileges to the hypervisor level and then delete the file in VM2. According to our previous work [5, 7], system calls constitute good forensic evidence, so a snapshot of VM2 captured during the attacking time was used to retrieve the system calls and kernel messages of the process "attack" and module "test". The system calls we obtained from the process of "attack" are shown in Figure 9, where the arguments following "execve" command in Line 1 clearly shows the attacker from VM1 used command line to execute an attack program named "attack", trying to delete the "samplefile.txt" located at the home folder of VM2(VM2 in named as "victim" in our experiment), and Line 25 clearly shows that the Linux module "test.ko" was injected to the Linux kernel of VM1 to do some work. The kernel activities of VM2 are illustrated in Figure 10, where the kernel messages between Line 1 and Line 6 show that the attacker from VM1 wrote some bytes to the memory after the address "ffff88007c723008", and the messages between Line 8 and Line 19 show that the attacker controlled the page table in Xen to execute the attacker's shellcode by linking the physical memory address where the shellcode is held to the virtual memory address in the page-table. This clearly shows the attacker used the shared memory to launch the attack.

```
1.  execve("./attack", ["./attack", "rm victim ~/samplefile.txt"], [/* 30 vars */]) = 0
2.  brk(NULL) = 0x8cd000
3. mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1,
0) = 0x7fa3a3022000
4. access("/etc/ld.so.preload", R_OK) = -1 ENOENT (No such file or directory)
5. open("/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3
…
25. open("test.ko", O_RDONLY) = 3
26. finit_module(3, "user_shellcmd_addr=1407334317317"..., 0) = 0
27. fstat(1, {st_mode=S_IFCHR|0620, st_rdev=makedev(136, 0), ...}) = 0
28. mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -
1, 0) = 0x7fa3a3021000
29. mmap(0x600000000000, 4096, PROT_READ|PROT_WRITE,
MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS|MAP_LOCKED, -1, 0) = 0x600000000000
30. delete_module("test", O_NONBLOCK) = 0
31. exit_group(0) = ?
```

Figure 9. The system calls obtained by tracing "attack" program

```
…
    1.  [  127.408066] write_byte_hyper(ffff88007c723008, 0x7)
    2.  [  127.436071] write_byte_hyper(ffff88007c723009, 0x90)
    3.  [  127.460074] write_byte_hyper(ffff88007c72300a, 0xba)
    4.  [  127.484055] write_byte_hyper(ffff88007c72300b, 0x26)
    5.  [  127.512054] write_byte_hyper(ffff88007c72300c, 0x1)
    6.  [  127.548083] write_byte_hyper(ffff88007c72300d, 0x0)
        …
    7.  [  127.628071] write_byte_hyper(ffff88007c723010, 0x0)
        …
    8.  [  127.660074] going to link PMD into target PUD
    9.  [  127.668058] linked PMD into target PUD
    10. [  127.676046] going to unlink mapping via userspace PUD
    11. [  127.684077] mapping unlink done
    12. [  127.692076] copying HV and user shellcode...
    13. [  127.700077] copied HV and user shellcode
    14. [  127.708066] int 0x85 returned 0x7331
    15. [  127.716077]   remapping paddr 0x21e8dd000 to vaddr 0xffff880079846800
    16. [  127.724076] IDT entry for 0x80 should be at 0xffff83021e8dd800
    17. [  127.732080] remapped IDT entry for 0x80 to 0xffff804000100800
    18. [  127.740077] IDT entry for 0x80: addr=0xffff82d080229ef0, selector=0xe008, ist=0x0,
        p=1, dpl=3, s=0, type=15
    19. [  127.748085] int 0x85 returned 0x1337
```

Figure 10. The kernel message from the injected module

Identifying an attack component is not a trivial task due to the nature of APTs. It requires a lot of detailed analysis such as looking at all processes and process threads that would have altered the state of an object, even under enhanced super-user privileges. As shown in our example, some of these missing steps may need to use system call logs. ATT&CK has techniques and tactics based on real-world observations, which makes it very helpful in identifying the corresponding processes or system calls that are related to a sub-attack phase of an APT attack.

4.2 Mapping Log Entries to Attack Steps

Once we identify the evidence by leveraging ATT&CK, we use cyber kill chain to map the corresponding evidence to different attack phases in order to construct the attack steps.

The evidence in Figure 5, Figure 6 and Figure 7 belong to the SQL injection attack, since the timestamps and alerts of these date are consistent. As our analysis in Section 4.1, the data in Figure 5 and 6 show that the attacker from 129.174.124.122 accessed the webserver (129.174.124.35) with SQL injection attempt("—1=1"), which is considered as "initial access" in ATT&CK, that can be easily mapped to the "weaponization" in pre-attack phase. The data in Figure 7 shows, at the same time the database had been queried by using the statement "select * from profiles where name='Alice' AND password='alice' or '1'='1'", which is clearly a SQL injection attack on the database. Since the database itself did not have any security mechanism, this implies the attack succeeded. Therefore, the attack had been delivered, and we can map the data in Figure 7 to "attack" phase in the cyber kill chain.

Likewise, we group the forensic data in Figure 8, Figure 9 and Figure 10 to the same attack by matching the process name "attack" and injected module name "test.ko". Because the data in Figure 8 and 9 only shows the attacker from VM1 ran the process "attack" and module "test" to do some work that did not show details, we map it to the "weaponization" phase that belongs to pre-attack stage. In addition, as we described in Section 4.1, the data in Figure 9 and 10 shows that the attacker manipulated the shared memory in the same hypervisor to execute shellcode on the victim VM, which can be mapped to "exploitation" phase that belongs to attack stage. Because we observed that the "samplefile.txt" file in the victim VM had been deleted, we knew the attack succeeded, which can be mapped to "action on objectives" in the cyber kill chain that belongs to the post-attack stage.

4.3 Co-relate Attack Steps to APTs (Cyber Kill Chains)

Our previous work [18] used a prolog-based tool to generate attack steps by using evidence in the form of Prolog predicates to instantiate rules composed of these predicates representing pre-conditions and post-conditions of an attack. These rules simulate generic attack techniques, which are written in the form of p :- p_1 , p_2 , $p_{n,}$ where predicate p represents the post-conditions of an attack, and predicates $p_{1,}$ p_2 ···, p_n represent the pre-conditions of the attack. The post-conditions refer to the privileges the attacker obtained after an attack, and the pre-conditions include the attacker's initial privilege, location, system configuration and the vulnerability used to launch the attack.

While this prolog-based tool can be used to generate attack steps, it requires users to categorize evidence to post-attack conditions and pre-attack conditions. The tool does not have predicates mapping to the seven cyber kill chain phases. Neither does it have corresponding rules that correlate the evidence corresponding to the seven phases of the cyber kill chain to pre-attack conditions and post-conditions. In this work, we propose to add the missing link by making changes as follows.

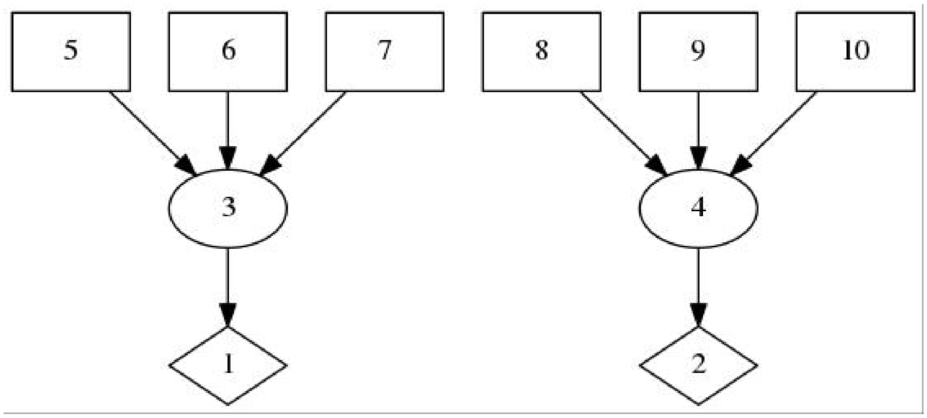
(1) We use predicate Pr_r, Pr_w, Pr_d, A_e, A_i, Po_c, Po_a to represent the pre-attack "reconnaissance", "weaponization", "delivery" phases, the attack "exploitation", "installation" phases, and the post-attack "command and control(C2)", "actions on objectives" phases respectively.

(2) We convert techniques in the ATT&CK matrix to corresponding predicates, and map them to the cyber kill chain phases as follows: (a) Predicates of “initial access” are mapped to Pr_r, (b) Predicates of “execution”, “persistence”, “privilege escalation”, “defense evasion”, “credential access” are mapped to Pr_w, (c) Predicates of “discovery” are mapped to A_i, (d) Predicates of “lateral movement” are mapped to A_i, (e) Predicates of “command and control” are mapped to Po_c, and (f) Predicates of “collection”, “exfiltration”, “impact” are mapped to $Po_{a.}$ Notice that these steps can be modeled as pre-conditions and post-conditions as given in Figure 11.

All predicates are composed of names and variables that depict the system configuration, attacker’s privilege, the network topology, software vulnerability and etc. For example, technique “exploit public-facing application” in “initial access” of the ATT&CK matrix is written to Predicate “Pr_r (attackerAccess(_host, _program))”, and technique “account manipulation” in “credential access” of the ATT&CK matrix is written to Predicate “Pr $_w$(hasAccount(_principal, _host, _account))”, where the variables (such as _host, _program …) following the predicate names(publicApp and hasAccount) in the predicates will be instantiated by concrete information during the run time of using this Prolog tool.

(3) We add rules to use cyber kill chain to correlate the corresponding predicates in different phases to an attack step. The rules are in the in the form of Po_c :- $(Pr_r\ ; Pr_w\ ; Pr_d)$, $(A_e\ ; A_i)$. *and* Po_a :- $(Pr_r\ ; Pr_w\ ; Pr_d)$, $(A_e\ ; A_i)$, where “;” represents logical OR, and “,” presents logical AND. These rules mean that, if there is evidence in pre-attack, attack and attack phases, an attack step is constructed.

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda. ”Forensic Analysis of Advanced Persistent Threat Attacks in Cloud Environments.”
Paper presented at Sixteenth IFIP 11.9 International Conference on Digital Forensics, New Delhi, IN. January 06, 2020 - January 08, 2020.



1	ExecCode(VM2,read)	6	networkServiceInfo(dbServer, httpd, tcp, 3660, user)
2	ExecCode(VM2,modify)	7	vulExists(webServer, 'CWE89', httpd)
3	THROUGH 3 (remote exploit of a server program)	8	hasAccount(attacker,VM1,root)
4	Through 8 (Access a host through executing code on the machine)	9	vulExists(sharedMemory,'CVE-2017-7228', hardware)
5	attackerAccess(publicWebApp)	10	vulProperty('CVE-2017-7228', remoteExploit, privEscalation)

Figure 11. The constructed attack steps of the experimental network

With these changes in our Prolog based tool, we constructed our attack steps as shown in Figure 11, which shows intuitive graphical attack steps for the two attacks constructed by using cyber kill chain to correlate evidence found in the attacked system. The left path shows that the attacker used a publicly available web application to launch SQL injection to the database in VM2, and the right path shows the attacker used the vulnerability in the shared hardware to attack VM2 and deleted a file in VM2.

We notice that our miniature example provides an initial example of the ATT&CK rule model. Although some rules (for example lateral movement by the attacker or pass the hash attacks) are missing. We postulate that these missing rules can be learned by using machine learning algorithms and incorporated into the steps of the ATT&CK process.

5. Conclusion

Justifying pre-attack, attack, and attack phases, requires having evidence of activities related to these phases. In APT attack analysis, there are many difficulties in constructing attack steps because (1) they do not lend themselves to traditional precondition, post-condition analysis (2) recognizing the pre-attack post-attack phases may require statistical correlation techniques that uses data from multiple sources. Thus, creating valid arguments for such attacks becomes more challenging. In particular assigning a "timestamp" for such an attack in a cloud environment becomes a challenge. We showed that we can leverage ATT&CK to identify the evidence and build the attack steps by mapping the evidence to different phases in the cyber kill chain. We used a case study to validate our framework. We plan to extend the relationship between the kill chain, evidence gathering and attack-attribution in the future work.

Disclaimer

This paper is not subject to copyright in the United States. Commercial products are identified in order to adequately specify certain procedures. In no case does such an identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

References:

1. S. M Milajerdi, R. Gjomemo, B. Eshete, R. Sekar and V.N. Venkatakrishnan, "HOLMES: real-time APT detection through correlation of suspicious information flows," *arXiv preprint arXiv:1810.01594*, 2018.
2. C. Liu, A. Singhal, D. Wijesekera, "A Layered Graphical Model for Cloud Forensic Mission Attack Impact Analysis", IFIP International Conference Digital Forensics, New Delhi, India, Jan 2018.
3. B.D. Bryant and H. Saiedian, "A novel kill-chain framework for remote security log analysis with SIEM software", *computers & security*, *67*, pp.198-210, 2017.

4. K. Ruan, J. Carthy, T. Kechadi,and M. Crosbie, "Cloud forensics", In IFIP International Conference on Digital Forensics, pp. 35-46. Springer Berlin Heidelberg, 2011.
5. C. Liu, A. Singhal and D. Wijesekera, "Identifying evidence for cloud forensic analysis", In IFIP International Conference on Digital Forensics (pp. 111-130). Springer, Cham, 2017.
6. http://libvmi.com/.
7. C. Liu, A. Singhal, R. Chandramouli and D. Wijesekera, "Determining forensic data requirements for detecting hypervisor attacks", International Conference on Digital Forensics, Orlando, U.S.A, Jan 2019.
8. Lockheed Martin, "Gaining the advantage--Applying Cyber Kill Chain® Methodology to Network Defense", retrieved from https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/Gaining_the_Advantage_Cyber_Kill_Chain.pdf.
9. C. Liu, A. Singhal, D. Wijesekera, "Relating admissibility standards for digital evidence to attack scenario reconstruction", Journal of Digital Forensics, Security and Law, 9(2) 15, 2014.
10. P. Gary (2001), 'A Road Map for Digital Forensic Research', Technical Report DTR-T001-01, DFRWS, Report from the First Digital Forensic Research Workshop (DFRWS).
11. MITRE ATT&CK™, retrieved from https://attack.mitre.org.
12. A. Pichan, M. Lazarescu and S. Soh, Cloud forensics: Technical challenges, solutions and comparative analysis, Digital Investigation, vol. 13, pp. 38–57, 2015.
13. J. Dykstra and A. Sherman, Acquiring forensic evidence from infrastructure-as-a-service cloud computing: Exploring and evaluating tools, trust and techniques, Digital Investigation, vol. 9(S), pp. S90–S98, 2012.
14. J. Dykstra and A. Sherman, Design and implementation of FROST: Digital forensic tools for the OpenStack cloud computing platform, Digital Investigation, vol. 10(S), pp. S87–S95, 2013.
15. S. Zawoad and R. Hasan, A trustworthy cloud forensics environment, in Advances in Digital Forensics XI, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 271–285, 2015.
16. B. Hay and K. Nance, Forensic examination of volatile system data using virtual introspection, ACM SIGOPS Operating Systems Review, vol. 42(3), pp. 74–82, 2008.
17. B. Dolan-Gavitt, B. Payne, W. Lee, "Leveraging forensic tools for virtual machine introspection", (Georgia Institute of Technology, Atlanta, GA), Technical Report, 2011.
18. C. Liu, A. Singhal and D. Wijesekera, "A probabilistic network forensic model for evidence analysis", In IFIP International Conference on Digital Forensics (pp. 189-210), Jan, 2016.
19. B. Strom, J. Battaglia, M. Kemmerer, W. Kupersanin, D. Miller, C. Wampler, S. Whitley and R. Wolf, "Finding cyber threats with ATT&CK-based analytics", Technical Report MTR170202, MITRE, 2017.
20. A. D'Amico and K. Whitley, "The real work of computer network defense analysts", In VizSEC 2007 (pp. 19-37).
21. K. Kent, S. Chevalier, T. Grance, "NIST SP 800-86 on Digital Forensics".

NIST SP 1285
September 2022

A Framework for Identifying and Prioritizing Data Analytics Opportunities in Additive Manufacturing

Hyunseop Park
Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Republic of Korea
hyunseop.park@postech.ac.kr

Hyunwoong Ko
School of Mechanical and Aerospace Engineering
Nanyang Technological University,
Singapore
kohy0060@e.ntu.edu.sg

Yung-Tsun T. Lee
Engineering Laboratory
National Institute of Standards and Technogy
Gaithersburg, MD
United States of America
yung-tsun.lee@nist.gov

Hyunbo Cho
Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Republic of Korea
hcho@postech.ac.kr

Paul Witherell
Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD
United States of America
paul.witherell@nist.gov

***Abstract*— Many industries, including manufacturing, are adopting data analytics (DA) in making decisions to improve quality, cost, and on-time delivery. In recent years, more research and development efforts have applied DA to additive manufacturing (AM) decision-making problems such as part design and process planning. Though there are many AM decision-making problems, not all benefit greatly from DA. This may be due to insufficient AM data, unreliable data quality, or the fact that DA is not cost effective when it is applied to some AM problems. This paper proposes a framework to investigate DA opportunities in a manufacturing operation, specifically AM. The proposed framework identifies and prioritizes AM potential opportunities where DA can make impact. The proposed framework is presented in a five-tier architecture, including value, decision-making, data analytics, data, and data source tiers. A case study is developed to illustrate how the proposed framework identifies DA opportunities in AM.**

Keywords—Data analytics, opportunity identification and prioritization, architecture, additive manufacturing

I. Introduction

Additive Manufacturing (AM) is a set of manufacturing technologies that join materials to produce three-dimensional (3D) objects from 3D solid models in layer-upon-layer ways opposed to the traditional subtractive manufacturing [1]. The tool-less and layer-upon-layer nature of AM provides unique capabilities of shape complexity, material complexity, hierarchical complexity, and functional complexity [2]. Such AM-enabled capabilities have largely lessened manufacturing constraints and significantly broadened design freedom, which offers new opportunities for developing functionally enhanced customized products [3]. AM is expected to lead the next generation of the manufacturing industry, the 1-batch customization era. To reach this expectation, it is important to improve performance in AM processes that eventually contributes to achieving the objectives of AM on first-part-correct and lead-time reduction.

Data analytics (DA) tools are expected to analyze data and produce actionable intelligence for the decision-makers. In recent years, the technology of DA has rapidly advanced [4]. Many researchers have demonstrated DA can help solve various manufacturing problems [5]. In this context, AM has been generating increasingly available data in the sense of volume, variety, and velocity [6]. The AM big data is providing great opportunities to use DA technologies and leverage DA capabilities to improve AM decision-making. Indeed, DA has been attracting attention in AM for data-driven decision-making [7].

DA studies in AM are still in the early stage. The majority of the existing DA studies in AM focuses on data analysis supporting only a few typical decision-making phases such as melt pool analysis for in-situ process signature monitoring [8]–[10]. This is because it is difficult to systemically map the decision-making and value chains to available DA capabilities and data, especially in AM where well-structured guidelines lack compared to other traditional manufacturing processes. It is necessary to systemically identify and prioritize DA opportunities in a complete view of AM decision-making that improves AM in general.

This paper proposes a novel framework for identifying and prioritizing DA opportunities in AM. The proposed framework has a five-tier architecture that consists of value, decision-making, data analytics, data, and data source tiers. Based on the architecture, the framework enables (1) identifying DA opportunities in AM and (2) prioritizing the identified DA opportunities. At the former phase, DA opportunities are identified with a top-down approach. The latter phase evaluates importance and feasibility (data readiness) of each identified DA opportunity.

The remainder of the paper is organized as follows. The second section describes background of AM data and DA opportunities. The third section introduces a five-tier architecture for AM data analytics that forms the foundation of the proposed framework. The fourth section presents the proposed framework for identifying and prioritizing DA opportunities in AM using the five-tier architecture. A case study in the fifth section implements the proposed framework. The paper is concluded with a brief conclusion and future work.

II. Background

A. AM Data

Advancements in sensor technologies have led to an unprecedented increase in AM data, encompassing many of

Funding agency: National Institute of Standards and Technology

Park, Hyunseop; Ko, Hyunwoong; Lee, Yung-Tsun; Cho, Hyunbo; Witherell, Paul. ”A Framework for Identifying and Prioritizing Data Analytics Opportunities in Additive Manufacturing.” Paper presented at 5th International Workshop on Methodologies to Improve Managing Big Data projects in IEEE big data 2019, Los Angeles, CA, US. December 09, 2019 - December 12, 2019.

the aspects of big data. From the volume perspective, AM generates terabyte (TB) size of in-situ monitoring data per build and TBs of computed tomography (CT) scan data [7]. From the variety perspective, AM generates data types including numerical data (e.g., machine logs), 2D images (e.g., thermal, optical), 3D (e.g., CAD models, CT scans), Audio (e.g., acoustic signals), and videos (e.g., thermal, optical) [7]. From the velocity perspective, up to 600 variables may be logged per second during the build [7]. The examples of each AM data are categorized and listed in Section III.

To capture, store, and properly manage for AM data, [11] proposes an additive manufacturing integrated data model (AMIDM) based on a product lifecycle management data modeling methods. Reference [12] presents a collaborative AM data management system, which is set to evolve through sharing of both the AM schema and AM development data among the stakeholders in the AM community. Reference [13] presents a digital thread and data package for AM to address not only data manageability but also traceability and accountability. As noted with the above studies, AM data management is studied actively but there are only few cases of DA using AM data.

B. *DA Opportunity Identification and Prioritization*

A DA opportunity can be defined as “a set of circumstances that makes DA possible to support and make impact on a decision-making issue”. Due to the lack of existing cases to refer, it is difficult to identify and prioritize DA opportunities in AM. Besides, there is no systematic method available to discover a DA opportunity. To address this issue, this paper proposes a two-phase approach: (1) identifying opportunities, and (2) prioritizing the identified opportunities. The methods for each phase used currently are reviewed as follows.

For opportunity identification, DA architectures that help describe each DA opportunity can be leveraged. Two popular architectures are used to describe DA. The Data, Information, Knowledge, Wisdom (DIKW) hierarchy uses four main components such as data, information, knowledge, and wisdom to describe DA [14]. However, the DIKW hierarchy has been criticized for its hard-to-consent definition of each tier [15]. The Analytics Canvas uses four-layer model including analytics use cases, data analysis, data pool, and data sources to describe DA use cases and the necessary data infrastructure [16]. However, the Analytics Canvas does not have a component that describes the value that each DA opportunity pursue. The DA opportunities may differ depending on the values pursued even in the same use case. For example, in the use case of predictive maintenance, the DA opportunity solves different problems depending on whether the value pursued is quality (good parts), performance (as fast as possible), or availability (no stop time).

For opportunity prioritization, multi-criteria decision-making methods, such as Fuzzy analytic hierarchy process (AHP), Fuzzy technique for order of preference by similarity to ideal solution (TOPSIS), and Fuzzy quality function deployment (QFD), can be used. Reference [17] uses Fuzzy AHP and Fuzzy TOPSIS to prioritize newly identified business model alternatives. Reference [18] uses Fuzzy QFD and Data Envelopment Analysis (DEA) to prioritize project portfolio. Reference [19] uses Fuzzy AHP and Fuzzy TOPSIS to prioritize solutions for reverse logistics barriers. Reference [20] uses Fuzzy AHP to select sustainable energy conversion technologies for agricultural residues.

III. Five-Tier Architecture for AM Data Analytics

A five-tier architecture for AM data analytics is presented in Fig. 1. The architecture is composed of the following tiers: (1) ‘Value Tier’ where values pursued in AM lifecycle are defined, (2) ‘Decision-Making Tier’ where AM decision-making activities are defined, (3) ‘Data-Analytics Tier’ where DA problems are defined, (4) ‘Data Tier’ where AM data and information are defined, and (5) ‘Data-Source Tier’ where AM data sources are defined. DA opportunity can be represented as a package composed of these five tiers.

There are relationships that drive the interactions between each tier. ‘Value Tier’ gives motivation to ‘Decision-Making

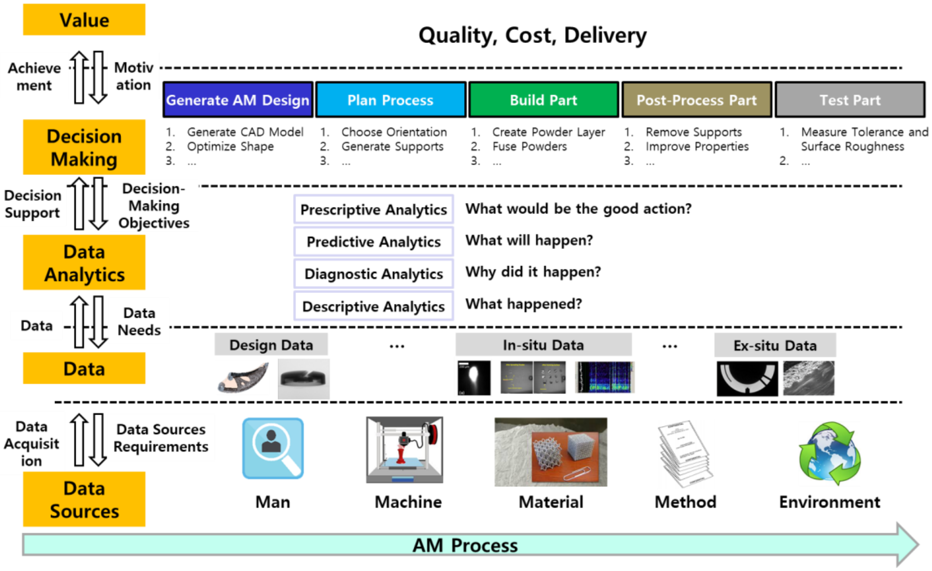


Fig. 1. 5-tier architecture for AM data analytics

Tier', 'Decision-Making Tier' achieves 'Value Tier'. 'Decision-Making Tier' gives decision making objectives to 'Data-Analytics Tier', 'Data-Analytics Tier' supports 'Decision-Making Tier'. 'Data-Analytics Tier' gives data requirements to 'Data Tier', 'Data Tier' provides data to 'Data-Analytics Tier'. Finally, 'Data Tier' gives data source requirements to 'Data-Source Tier' and 'Data-Source Tier' provides data sources to 'Data Tier'. Each tier is explained in more detail as follows.

A. Value Tier

At the Value Tier, values pursued in AM lifecycle are defined in terms of Quality, Cost, and Delivery (QCD) [21]. QCD is about value dimensions used to assess production process. Quality in AM can be performance, conformance, durability, and aesthetics. Cost in AM can be labor costs, material costs, energy costs, and maintenance costs. Delivery in AM can be on-time delivery, right amount delivery, and right location delivery. Examples of related research works in AM are in improving quality [22], reducing cost [23], and realizing on-time delivery [24].

B. Decision-Making Tier

At the Decision-Making Tier, decision-making activities for AM are defined. These activities follow the design-to-product transformation of the part, beginning at early design stages and ending with a finished part. The predefined activity model for Laser Powder Bed Fusion (LPBF) process [13], [25], which is represented using the IDEF0 [26], is used as an example. The top three levels of the decision-making activities listed in that model are shown in Table I. It includes six of Level-1 activities: 'Generate AM Design', 'Plan Process (Machine Independent)', 'Plan Process (Machine Dependent)', 'Build Part', 'Post-process Part', and 'Test Part'. These six activities are decomposed to twenty-two (22) Level-2 and thirty-four (34) Level-3 sub-activities. Decision-making objectives for each decision-making activity are also defined at this tier.

Using IDEF0, the decision-making activities can be defined by functions and related components such as input (I), control (C), output (O), and mechanism (M), or ICOM [26]. A function describes what an activity should accomplish. An input can be data, objects, or materials that are transformed by the activity. A control can be one or more conditions necessary for the activity to create correct outputs. An output is a set of results generated by the activity. A mechanism is the means that support the execution of the activity such as software, equipment, and personnel. An example is shown in Fig.2 [25]. The activity 'A11: Generate CAD Model' itself is a function. The input of the activity is a design, which refers to the conceptual design of a part and related design requirements. The control of the activity is guidelines, which refers to the design guidance that may be provided from feedback opportunities or part of governing design or process or material requirements. It is at the controller where many of the decision-making opportunities can be identified, as the configuration of the controller will impact the output of the activity. Here, the output of the activity is a CAD model, which refers to the computer-generated geometry that was developed using a CAD software. The mechanism of the activity is software, which is used to transform the design into a CAD model.

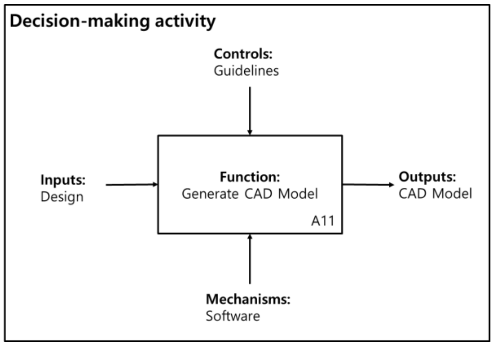


Fig. 2. Representation of a decision-making activity [25]

TABLE I. HIGH LEVEL DECISION-MAKING ACTIVITIES IN AM

Level-1	*Level-2*	*Level-3*
A1: Generate AM Design	A11: Generate CAD Model	-
	A12: Optimize Shape	-
	A13: Tessellate Model	-
	A14: Repair Tessellated Model	-
	A15: Modify Tessellated Model	-
A2: Plan Process (Machine Independent)	A21: Choose Orientation	-
	A22: Generate Supports	-
A3: Plan Process (Machine Dependent)	A31: Setup Tessellated Model	A311: Determine Orientation
		A312: Design Supports
	A32: Create Build Model	A321: Place Part
		A322: Generate Slices
		A323: Generate Scan Strategy
	A33: Plan Powder Fusion Strategy	A331: Set Quality Parameters
		A332: Set Control Parameters
		A333: Set Powder Fusion Parameters
		A334: Set Recoating Parameters
	A34: Plan Monitoring Strategy	-
A4: Build Part	A41: Create Powder Layer	-
	A42: Fuse Powders	-
	A43: Monitor Fusion	-
A5: Post-process Part	A51: Remove Supports	-
	A52: Improve Properties	-
	A53: Finish Part	-
A6: Test Part	A61: Measure Tolerances and	A611: Measure External Tolerances

Level-1	*Level-2*	*Level-3*
	Surface Roughness	A612: Measure Internal Tolerances
		A613: Measure Surface Roughness
	A62: Measure Porosity and Cracks	A621: Measure Part Porosity
		A622: Identifying and Measure Cracks
	A63: Measure Part Properties	A631: Measure Mechanical Properties
		A632: Measure Microstructures Properties
		A633: Measure Electrical Properties
		A634: Measure Chemical Properties
		A635: Measure Thermal Properties
	A64: Evaluate Test Results	-

C. Data-Analytics Tier

At the Data-Analytics Tier, each decision-making objective may be translated in DA perspectives with the following four types of analytics: prescriptive analytics, predictive analytics, diagnostic analytics, and descriptive analytics [16]. Each type of analytics is described as follows. Prescriptive analytics is to answer the question of what action should be done. Prescribing optimized powder fusion parameters is an example of prescriptive analytics in AM. Predictive analytics is to answer when and what will happen. Predicting porosity is an example of predictive analytics in AM. Diagnostic analytics is to answer the question of why it happened. Identifying the relationship between design parameters and surface roughness is an example of diagnostic analytics in AM. Descriptive analytics is to answer the question of what happened. Characterizing melt pool behavior is an example of descriptive analytics.

As the type of DA is defined, it is often easier to choose suitable algorithms in a practical use. For prescriptive analytics, reinforcement learning algorithms, such as Deep Q-Learning, and recommender system algorithms, such as associate rule mining, can be used to find best action or optimize problems [27]. For predictive analytics, supervised learning algorithms, such as neural networks and support vector machine, can be used to develop predictive model [27]. For diagnostic analytics, unsupervised learning algorithms, such as K-means clustering, can be used for grouping a set of objects; and supervised learning, such as linear regression, can be used to identify causal relationships [28]–[30]. For descriptive analytics, general statistics, like mean, min, and max; and signal processing algorithms, like Wavelet Transform and Fast Fourier Transform, can be used to obtain meaningful descriptive information from raw data [31].

D. Data Tier

At the Data Tier, AM data types are defined. AM generates a variety of data. Examples are listed as follows [7].

- Material properties: material chemistry, material microstructure, powder size distribution, etc.
- Design parameters: wall thickness, orientation, overhang angle, etc.
- Process parameters: laser power, scan speed, hatch spacing, machine logs, etc.
- Process signatures: thermal data (e.g., melt pool width, temperature), optical images and videos, acoustic signals, etc.
- Part property: tensile toughness, hardness, etc.
- Product performance: fatigue life, corrosion, meets design criteria, etc.

AM data can be stored in database systems. The Additive Manufacturing Materials Database (AMMD) [32] built by National Institute of Standards and Technology (NIST) is an example.

E. Data-Sources Tier

An AM lifecycle is composed of five stages: (1) Design, (2) Process Plan, (3) Build, (4) Post Process, and (5) Test and Validation [11]. Each stage produces data from its data sources. In this tier, the data sources can be categorized into Man, Machine, Material, Method, and Environment (4M1E) [33].

4M1E are the foundation resources managed for QCD in production systems. To achieve QCD, data from 4M1E needs to be analyzed. Among 4M1E, Man means participants in the AM lifecycle such as part designers and process planners. Machine can be every machine used in the AM lifecycle including coordinate measuring machine and AM machine, also known as 3D printer. Material can be any material used in the AM lifecycle such as plastic, metal powder, or ceramic. Method in the AM lifecycle can be AM standards, part measurement methods, test specifications, etc. Finally, Environment in the AM lifecycle can be software, workplace, temperature, energy usage, etc.

IV. Proposed Framework

The proposed framework uses a five-tier architecture with sequential steps. By applying the proposed framework to set an overall DA -direction in AM, we use a two-phase approach. First, the proposed framework helps to identify DA opportunities through a top-down approach. Second, the proposed framework helps prioritize the identified DA opportunities in terms of importance and feasibility. Here, the feasibility refers to data readiness. The detail processes are described as follows.

A. Phase 1 - Identifying DA Opportunities

DA opportunities are identified with a top-down approach, as shown in Fig. 3. First, AM-lifecycle values are determined using QCD at the Value Tier. The value(s) can be one or more of the following: 'Quality', 'Cost', and 'Delivery', or their extensions, such as "Aesthetics" and "labor costs efficiency".

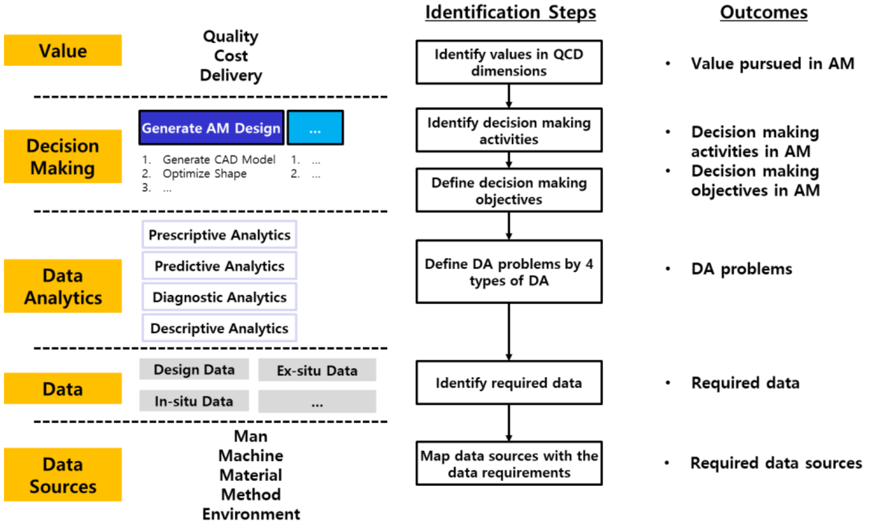


Fig. 3. Identification of DA Opportunities

Decision-making activities and decision-making objectives are identified in the 'Decision-Making Tier'. Decision-making activities are related to a pre-defined value from the Value Tier, here identified from the existing activity models, e.g., Table I. As mentioned in Section III, the decision-making activities can be broken into a set of functions and ICOMs. Once the value and decision-making activities are defined, a decision-making objective can be stated as "**Improving + [value] + when + [decision-making activity]**", where "Improving" and "when" provide syntax. For example, if the target value and decision-making activity are defined as "Material cost efficiency" and "A11: Generate CAD Model", then the corresponding decision-making objective can be stated as "Improving material cost efficiency when Generate CAD Model".

Once a decision-making objective is defined, potential types of DA problems are defined for each decision-making objective at the 'Data-Analytics Tier'. The syntax of each DA problem is summarized as follows. Note that these syntaxes are examples and are for the illustration purpose. A more complete syntax is currently under development.

- Prescriptive analytics: **"Prescribing + [C] + to maximize + [V]"**
- Predictive analytics: **"Predicting + [V] + based on given + [ICOM]"**
- Diagnostic analytics: **"Identifying relationship between + [ICOM] + and + [V] + based on their characteristics"**
- Descriptive analytics: **"Characterizing + [I] + [C] + [O] + [M] + [V]"**

Where V represents a predefined value from the Value Tier, and I is a input, C is a control, O is an output, M is a mechanism defined in the ICOM.

The goal of the prescriptive analytics is to directly support the achievement of the objectives of the corresponding decision-making activity in the IDEF0. Therefore, the prescriptive analytics tier in the framework identifies prescriptive analytics problems that consider DA objectives based on V, which aims to construct data-driven prescriptive guidelines. Based on the structured syntax of prescriptive analytics problems, following prescriptive analytics supports the development of a condition in C that can maximize V. The syntax is intended to lead associated prescriptive analytics to specifically support a transformation of an input I to a desired O with a consideration of M identified from the IDEF0 representation. Based on the syntax, prescriptive analytics can be then designed under possible situations or scenarios, to suggest courses of actions or strategies.

The predictive analytics tier aims to identify analytics problems to predict the target V. V is intended to be maximized in the prescriptive analytics. Therefore, V is used as criteria for suggesting courses of actions or strategies in the prescriptive analytics. To support this, the predictive analytics tier sets problems to predict information about possible situations or scenarios influencing V. Leveraging V and ICOM, the proposed syntax of predictive analytics problems helps to design data analytics that provides predictive criteria for associated prescriptive analytics.

The diagnostic analytics tier sets analytics problems for identifying the corelationships between ICOM and V in the past, or identifying which ICOM relationships influenced V in the past. The proposed syntax of diagnostic analytics problems helps to set data analytics approaches that provide bases the associated predictive analytics selects/extracts predictive parameters based upon. An example of the diagnostic analytics that can be considered is data analysis of ICOM relationships as root causes of the target V. Such analysis can be pursued leveraging the proposed syntax to support the V prediction of which analytics is designed in the upper-level predictive analytics tier.

The descriptive analytics tier identifies analytics problems to characterize ICOM and V to support the identification of relationships in the diagnostic analytics. It generates descriptive parameters representing data requirements that describe the required characteristics or behaviors of ICOM- and V-specific data. The ICOM- and V-specific data are then adaptively requested in the data tier to support the identification of DA opportunities.

The following examples illustrate the four DA problem types by using the same decision-making objective discussed previously, i.e., “Improving material cost efficiency when Generate CAD Model”. Note that the ICOM of A11 are I: Design, C: Design Guidelines, O: CAD Model, and M: CAD Software. To achieve that objective, a prescriptive analytics problem can be defined as “Prescribing Design Guidelines to maximize Material cost efficiency”. A predictive analytics problem can be “Predicting Material cost efficiency based on given Design, Design Guidelines, CAD Model, CAD Software”. A diagnostic analytics problem can be “Identifying relationship between Design, Design Guidelines, CAD Model, CAD Software and Material cost efficiency based on their characteristics”. Finally, a descriptive analytics problem can be “Characterizing each Design, Design Guidelines, CAD Model, CAD Software, and Material cost efficiency”.

Each defined DA problem, presented using the proposed syntax, with some grammar modifications when necessary, is treated as the title of a DA opportunity. Note that descriptive analytics can be sometimes substituted by the outputs of other decision-making activities. For example, when “Characterizing aesthetic”, other quality indicators such as surface roughness of ‘A613: Measure Surface Roughness’ can be used.

To realize the objectives of the DA problems data requirement for each DA problem should be defined at the ‘Data Tier’. For the above example, the descriptive analytics for “Characterizing each Design, Design Guidelines, CAD Model, CAD Software, and Material cost efficiency” requires data related to the Design, Design Guidelines, CAD Model, and Material cost efficiency. The prescriptive analytics for “Prescribing Design Guidelines to maximize Material cost efficiency” requires not only to have data related to the Design, Design Guidelines, and CAD Model, and Material cost efficiency but also the results of other analytics including the predicted Material cost efficiency as well. The data requirements of the DA problem will influence the feasibility of each DA opportunity. This is because the feasibility is mainly measured based on the data readiness level.

Finally, data sources are defined in the ‘Data-Source Tier’. For instance, to satisfy the data requirements for the “Characterizing Design, Design Guidelines, CAD Model, CAD Software, Material cost efficiency,” it requires data from these data sources: part designer, design guideline documents, CAD software, and cost measurement method.

Often, we identify many DA opportunities through the phase 1 process. In reality, it is difficult and also not necessary to research and develop all the identified DA opportunities, due to the constraints of time, cost, and importance. Therefore, prioritizing the DA opportunities is an important task.

B. Phase 2 - Prioritizing DA Opportunities

As shown in Fig. 4., the prioritization phase is broken into two parts: evaluating the importance of each DA opportunity and evaluating the feasibility of each DA opportunity. Decision makers (DMs) are the key persons to participate in the evaluation.

To evaluate the importance of each DA opportunity, the evaluation is performed by the DMs, starting from the ‘Value Tier’ to the ‘Decision-Making Tier’, and then to the ‘Data-Analytics Tier’. First, the importance of each identified value is assessed. Then, the importance of improving each identified value for the corresponding decision-making activity is evaluated and rated. At last, the importance of each DA problem for the decision-making activity is evaluated and rated. Finally, using a hierarchical multi-criteria decision-making method, the importance of each DA opportunity can be calculated.

To evaluate the feasibility of each DA opportunity, the evaluation is performed through ‘Data Source Tier’ to ‘Data Tier’ and ‘Data Tier’ to ‘Data-Analytics Tier’. First, the feasibility of data satisfying the requirements for each DA opportunity is investigated. Then, the importance of each data for each DA problem is evaluated. Finally, the data readiness

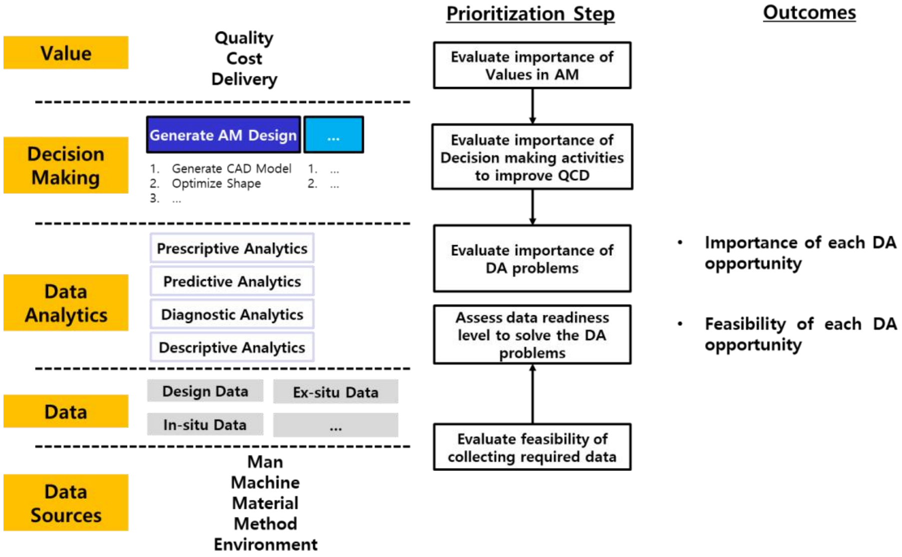


Fig. 4. Prioritization of DA opportunities

level for each DA opportunity is assessed by performing a gap analysis between the feasibility and the importance of the required data.

A sample output of the prioritization, i.e., Phase 2, using the proposed framework is shown in Fig. 4. The identified DA opportunities from Phase 1 are mapped into the prioritization matrix, where x-axis represents feasibility and y-axis represents importance. Based on the prioritization matrix, the identified DA opportunities can be classified into four groups, as shown in Fig. 5.

When mapping DA opportunities, the DA opportunities in the high importance and high feasibility groups are critical to work on. On the other hand, the DA opportunities in the low importance and low feasibility group are considered the lowest priority. The DA opportunities in the high importance and low feasibility group are potentially critical projects, where new solutions may be needed to improve the data readiness level and hence upgrade the prioritization group of the opportunity. The DA opportunities in the low importance and high feasibility group may be easy to develop but most likely the effort is not beneficial, however, sometimes such opportunities might be useful for proof of DA concept in emerging areas.

V. Case study

Phase 1 of the proposed framework is illustrated in this case study on LPBF processes. Table II shows the outputs of the proposed framework in the overall cases. Three cases, Case 1, Case 2, and Case 3 are identified by the proposed top-down approach.

The case study begins by taking the input Vs: surface texture quality, part porosity reducibility, and time efficiency. The Vs are mapped to associated decision-making activities of 'Design', 'Build Part', and 'Test Part' in Case 1, Case 2, and Case 3, respectively, in the Decision-making Tier. Then, the target levels of decision-making abstraction are systematically identified through the decomposition of the functional relationships in the IDEF0 that represents the Decision-making Tier. The AM activities identified at the target abstraction levels are "Generate Detailed Design Model for LPBF", "Monitoring In-situ Process Signatures of LPBF Behaviors", and "3D Scan Part". Once the target AM activity is determined, its corresponding ICOMs and decision-making objectives are also identified for the DA.

At the Data Analytics Tier, the target Vs, ICOMs, and decision-making objectives are leveraged to define DA objectives utilizing the proposed DA objective structures. For example, in Case 1, the Prescriptive Analytics has an objective of "Prescribing design rule to maximize surface texture quality". This objective is identified to prescriptively guide what specific design actions the activity A12-n should pursue to maximize surface texture quality for LPBF. The required information for this prescriptive analytics is the predicted surface texture quality (V for Case 1) based on based on given overhang design model, design rule, process plan, material properties, redesigned part. Based on this requirement, the Predictive Analytic formulates "Predicting surface texture quality based on given overhang design model, design rule, process plan, material properties, redesigned part".

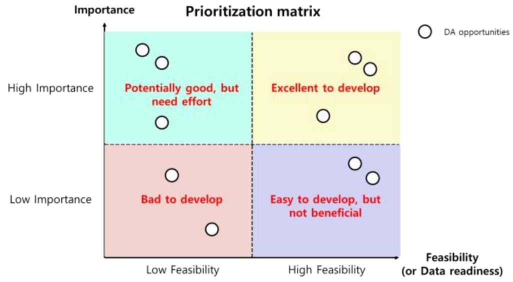


Fig. 5. Prioritization matrix

The Dignostic Analytics in Case 1 aims to solve an associated problem to support the predictive analytics. Case 1's predictive analytics requires identified relationships between the overhang design model (I for Case 1), design rule for overhang features, process plan, material properties (C for Case 1), and surface texture quality (V for Case 1). Based on this requirement, a diagnostic analytics objective is set as "Identifying relationship between overhang design model, design rule, process plan, material properties, and surface texture quality based on their characteristics". Such objectives help group data features and set necessary hypotheses in predictive analytics. Finally, the Descriptive Analytics has an objective of "Characterizing each overhang design model, design rule, process plan, material properties" to generate descriptive parameters that describe the characteristics and behaviors of Case 1-specific data.

In the proposed top-down approach, the DA objective structures enable selecting or requesting AM data sets suitable for each type of data analytics. At the same time, the DA objective structures become a bridge that links values and decision-making objectives to data requirements. Therefore, the top-down approach adaptively generates data requirements that are goal-oriented as well as DA type-specific. Such advantage provides varying DA opportunities even for the same data sources that bottom-up approaches may not be able to capture. The data of material type are good examples while they are commonly required in the 3 cases for different values, decision-making objectives, and DA types. When required data sets are not available, the top-down approach can generate a request for new AM data to incorporate data requirements into plans for further data obtainments [34]. Examples of the further data obtainment can be fusions of available data, field tests and simulations, and installations of sensor environments guided by the top-down approach.

The case study identified twelve DA opportunities from the three cases in the AM lifecycle. It is meaningful that the DA opportunities are identified without heavily referring to existing opportunities identified from other existing DA studies. This study is expected to serve as a structured guideline for researchers and practitioners who seek new DA opportunities in AM. However, the case study did not prioritize the identified DA opportunities yet. More opportunities remained to be identified so we will prioritize them in the future work.

TABLE II. AN OVERVIEW OF CASE STUDIES: OUTPUT OF EACH TIER

Tier	**Sub-items**	**Case Studies**		
		Case 1	***Case 2***	***Case 3***
Value	Quality	Surface texture quality	Part porosity reducibility	-
	Cost	-	-	-
	Delivery	-	-	Time efficiency
Decision Making	Decision Making Activity	A1: Design A12: Optimize shape A12-n: Generate Detailed Design Model for LPBF - I: Overhang design model - C: Design rule for overhang features, Process plan, Material properties - O: Redesigned part - M: CAD Software	A4: Build Part A43: Monitoring fusion A43-n: Monitoring in-situ process signatures of LPBF behaviors - I: In-situ process signatures - C: In-situ defect evaluation rule/guideline, Process plan, material properties - O: Evaluation results - M: In-situ monitoring cameras and software	A6: Test Part A61: Measure Tolerance and Surface Roughness A611: Measure External Tolerances A611-n: 3D Scan part - I: Target part - C: 3D Scan path, material properties - O: Reconstructed 3D model - M: Robotic 3D scanning system
	Decision Making Objective	Improving surface texture quality when generating detailed design model for LPBF	Improving part porosity reducibility when monitoring in-situ process signatures of LPBF behaviors	Improving time efficiency when 3D Scan part
Data Analytics	Prescriptive Analytics	Prescribing design rule to maximize surface texture quality	Prescribing in-situ defect evaluation rule/guideline to maximize part porosity reducibility	Prescribing 3D scan path to maximize time efficiency
	Predictive Analytics	Predicting surface texture quality based on given overhang design model, process plan, material properties	Predicting part porosity reducibility based on given in-situ process signatures, process plan, material properties	Predicting time efficiency based on given target part, 3D scan path, and material properties
	Diagnostic Analytics	Identifying relationship between overhang design model, process plan, material properties, and surface texture quality based on their characteristics	Identifying relationship between in-situ process signatures, process plan, material properties, and part porosity reducibility	Identifying relationship between target part, 3D scan path, material properties and time efficiency based on their characteristics
	Descriptive Analytics	Characterizing overhang design model, design rule, process plan, material properties	Characterizing in-situ process signatures, process plan, material properties, part porosity reducibility	Characterizing target part, 3D scan path, material properties, time efficiency
Data	For Prescriptive Analytics	Data of predictive surface roughness with given overhang geometry, process plan, and material properties changes	Data of predictive porosity density, shape, location with given in-situ process signatures, process plan, and material properties	Data of predictive scan time with given target part, scan path, reconstructed 3D model, material properties, scanning environment
	For Predictive Analytics	Historical data on the surface roughness as response variable Selected/extracted features from overhang geometry, process plan, and material properties as input variables	Historical data on the porosity density, shape, location as response variable Selected/extracted features from in-situ process signatures, process plan, and material properties as input variables	Historical data on the scan time as response variable Selected/extracted features from target part, scan path, reconstructed 3D model, material properties, and scanning environment as input variables
	For Diagnostic Analytics	Historical data on the surface roughness as response variable Historical data on overhang geometry, process plan, and material properties as input variable	Historical data on the porosity density, shape, location as response variable Historical data on in-situ process signatures, process plan, and material properties as input variable	Historical data on the scan time as response variable Historical data on target part, scan path, reconstructed 3D model, material properties, and scanning environment as input variables
	For Descriptive Analytics	Surface roughness data, Overhang geometry data (e.g. overhang type, downskin, surface angle, overhang dimension), Process plan data (e.g. tool path, energy source, power, speed, hatching distance, layer thickness, part position and orientation), and Material properties (e.g. density, powder distribution, material type), Design rule data	Porosity data (e.g. density, shape, location), In-situ process signatures (e.g. melt pool size, shape), Process plan data (e.g. tool path, energy source, power, speed, hatching distance, layer thickness, part position and orientation), and Material properties (e.g. density, powder distribution, material type), Design rule data	Scan time data, Target part data (e.g. CAD data), Scan path data, Reconstructed 3D model (e.g. point cloud), Material properties (e.g. material type), and Scanning environment data (e.g. light condition)
Data Source	Man	-	-	-
	Machine	LPBF machine, Ex-situ part measurement machine (e.g., XCT scanner)	LPBF machine, in-situ monitoring camera (e.g., optical camera), ex-situ part measurement machine (e.g., XCT scanner)	3D scanning robot
	Material	Process planning model, LPBFed part model	In-situ process signature model, process planning model, LPBFed part model	Target part
	Method	CAD development, LPBF machine control method, surface and volume measurement system (e. g., XCT scanner)	LPBF machine control method, in-situ monitoring method, surface and volume measurement system (e. g., XCT scanner)	3D reconstruction method
	Environment	-	-	Scanning environment, CAD software

VI. Conclusion

Although AM generates big data that provides opportunities to use DA, the AM community does not have many successful stories of the DA applications. The lack of DA cases makes it difficult for researchers and practitioners in AM to define the AM problems where DA can have an impact. To address this issue, this paper proposes a framework in a five-tier architecture, including value, decision-making, data analytics, data, and data sources, to help (1) identify and (2) prioritize DA opportunities in AM. For the former phase, a top-down approach in the five-tier architecture is formulated with a set of suggested syntaxes for describing potential DA opportunities. For the latter phase, there are two dimensions to be evaluated for prioritization: importance and feasibility. By using the proposed framework, a case study identified twelve DA opportunities in LPBF processes. These DA opportunities were identified from different values, decision-making activities, and DA types, where values are the targets the improvement aims at, decision-making activities are the AM activities that could potentially be improved in terms of quality, cost and delivery, and DA types include prescriptive, predictive, diagnostic, and descriptive analytics. The case study demonstrates the proposed framework could systematically identify potential DA opportunities in a complete view of the AM lifecycle, without heavily relying on the existing DA studies in AM. Since this framework does not stick to a certain domain so it is expected to contribute to not only AM but also other domain where DA is pre-mature.

The future work will focus on formalizations of the proposed framework that will enable the identification and prioritization of DA opportunities in a consistent way. We will continue to develop novel sets of syntax structures where the proposed formulations of DA objectives are expanded. The range of the syntax structure formulation will be expanded for the other tiers in the proposed architecture as well. The future work will then transform the prioritization phase into a formal method. The formal method will be equipped with the prioritization techniques mentioned in Section II. Furthermore, the formalized framework will be implemented in software environments with AM data and knowledge bases to automatically support identifying and prioritizing DA opportunities. The continuation of this study will eventually provide a set of DA opportunities with higher data readiness level and higher impact to the AM community.

Acknowledgment

The authors acknowledge the support by the Additive Manufacturing Program at the National Institute of Standards and Technology (NIST), US Department of Commerce. Certain commercial systems are identified in this article. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

References

[1] "ASTM F2792-12a, Standard Terminology for Additive Manufacturing Technologies (Withdrawn 2015)." ASTM International, West Conshohocken, PA, 2012.

[2] I. Gibson, D. Rosen, and B. Stucker, "Design for Additive Manufacturing," in *Additive Manufacturing Technologies: 3D Printing, Rapid Prototyping, and Direct Digital Manufacturing*, New York, NY: Springer New York, 2015, pp. 399–435.

[3] H. Ko, S. K. Moon, and J. Hwang, "Design for additive manufacturing in customized products," *Int. J. Precis. Eng. Manuf.*, vol. 16, no. 11, pp. 2369–2375, 2015.

[4] S. J. Qin and L. H. Chiang, "Advances and opportunities in machine learning for process data analytics," *Comput. Chem. Eng.*, vol. 126, pp. 465–473, 2019.

[5] T. Wuest, D. Weimer, C. Irgens, and K. D. Thoben, "Machine learning in manufacturing: Advantages, challenges, and applications," *Prod. Manuf. Res.*, vol. 4, no. 1, pp. 23–45, 2016.

[6] P. Witherell, "Emerging Datasets and Analytics Opportunities in Metals Additive Manufacturing," in *Direct Digital manufacturing Conference*.

[7] S. S. Razvi, S. Feng, A. Narayanan, Y.-T. T. Lee, and P. Witherell, "A review of machine learning applications in additive manufacturing," in *Proceeding of the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2019. in press.

[8] L. Scime and J. Beuth, "Using machine learning to identify in-situ melt pool signatures indicative of flaw formation in a laser powder bed fusion additive manufacturing process," *Addit. Manuf.*, vol. 25, no. November 2018, pp. 151–165, 2019.

[9] Z. Yang, L. Yan, H. Yeung, and S. Krishnamurty, "Investigation of Deep Learning for Real-Time Melt Pool Classification in Additive Manufacturing."

[10] Y. Zhang, G. S. Hong, D. Ye, K. Zhu, and J. Y. H. Fuh, "Extraction and evaluation of melt pool, plume and spatter information for powder-bed fusion AM process monitoring," *Mater. Des.*, vol. 156, pp. 458–469, 2018.

[11] Y. Lu, S. Choi, and P. Witherell, "Towards an integrated data schema design for additive manufacturing: Conceptual modeling," *Proc. ASME Des. Eng. Tech. Conf.*, vol. 1A-2015, no. August, 2015.

[12] Y. Lu, P. Witherell, and A. Donmez, "A collaborative data management system for additive manufacturing," *Proc. ASME Des. Eng. Tech. Conf.*, vol. 1, pp. 1–7, 2017.

[13] D. B. Kim, P. Witherell, Y. Lu, and S. Feng, "Toward a Digital Thread and Data Package for Metals-Additive Manufacturing," *Smart Sustain. Manuf. Syst.*, vol. 1, no. 1, p. 20160003, 2017.

[14] J. Gu and L. Zhang, "Data, DIKW, big data and data science," *Procedia Comput. Sci.*, vol. 31, pp. 814–821, 2014.

[15] M. Fricke, "The knowledge Pyramid: A Critique of the DIKW Hierarchy," *J. Inf. Sci.*, vol. 35, no. 2, pp. 1–13, 2007.

[16] A. Kühn, R. Joppen, F. Reinhart, D. Röltgen, S. Von Enzberg, and R. Dumitrescu, "Analytics Canvas - A Framework for the Design and Specification of Data Analytics Projects," *Procedia CIRP*, vol. 70, pp. 162–167, 2018.

[17] K. Im and H. Cho, "A systematic approach for developing a new business model using morphological analysis and integrated fuzzy approach," *Expert Syst. Appl.*, vol. 40, no. 11, pp. 4463–4477, 2013.

[18] H. Jafarzadeh, P. Akbari, and B. Abedin, "A methodology for project portfolio selection under criteria prioritisation, uncertainty and projects interdependency – combination of fuzzy QFD and DEA," *Expert Syst. Appl.*, vol. 110, pp. 237–249, 2018.

[19] P. Sirisawat and T. Kiatcharoenpol, "Fuzzy AHP-TOPSIS approaches to prioritizing solutions for reverse logistics barriers," *Comput. Ind. Eng.*, vol. 117, no. September 2017, pp. 303–318, 2018.

[20] B. Wang, J. Song, J. Ren, K. Li, H. Duan, and X. Wang, "Selecting sustainable energy conversion technologies for agricultural residues: A fuzzy AHP-VIKOR based prioritization from life cycle perspective," *Resour. Conserv. Recycl.*, vol. 142, no. October 2018, pp. 78–87, 2019.

[21] M. Imai, "Gemba Kaizen. A Commonsense, Low-Cost Approach to Management," Jan. 2007.

[22] P. Chesser *et al.*, "Extrusion control for high quality printing on Big Area Additive Manufacturing (BAAM)systems," *Addit. Manuf.*, vol. 28, no. September 2018, pp. 445–455, 2019.

[23] A. Mahadik and D. Masel, "Implementation of Additive Manufacturing Cost Estimation Tool (AMCET) Using Break-down Approach," *Procedia Manuf.*, vol. 17, pp. 70–77, 2018.

Park, Hyunseop; Ko, Hyunwoong; Lee, Yung-Tsun; Cho, Hyunbo; Witherell, Paul. "A Framework for Identifying and Prioritizing Data Analytics Opportunities in Additive Manufacturing." Paper presented at 5th International Workshop on Methodologies to Improve Managing Big Data projects in IEEE big data 2019, Los Angeles, CA, US. December 09, 2019 - December 12, 2019.

[24] D. Zindani and K. Kumar, “An Insight into Additive Manufacturing of Fiber Reinforced Polymer Composite,” *Int. J. Light. Mater. Manuf.*, 2019.

[25] R. Lipman, P. Witherell, S. Leong, and Y. Lu, “An Activity Model for Additive Manufacturing Powder Bed Fusion,” 2016. unpublished.

[26] N. I. of S. and Technology, “Integration Definition for Function Modeling (IDEF0),” 1993.

[27] K. Lepenioti, A. Bousdekis, D. Apostolou, and G. Mentzas, “Prescriptive analytics: Literature review and research challenges,” *Int. J. Inf. Manage.*, vol. 50, no. October 2018, pp. 57–70, 2020.

[28] H. Lorentz, O. P. Hilmola, J. Malmsten, and J. S. Srai, “Cluster analysis application for understanding SME manufacturing strategies,” *Expert Syst. Appl.*, vol. 66, pp. 176–188, 2016.

[29] E. Uhlmann, R. P. Pontes, C. Geisert, and E. Hohwieler, “Cluster identification of sensor data for predictive maintenance in a Selective Laser Melting machine tool,” *Procedia Manuf.*, vol. 24, pp. 60–65, 2018.

[30] I. Baturynska, “Statistical analysis of dimensional accuracy in additive manufacturing considering STL model properties,” *Int. J. Adv. Manuf. Technol.*, vol. 97, no. 5–8, pp. 2835–2849, 2018.

[31] S. A. Shevchik, C. Kenel, C. Leinenbach, and K. Wasmer, “Acoustic emission for in situ quality monitoring in additive manufacturing using spectral convolutional neural networks,” *Addit. Manuf.*, vol. 21, pp. 598–604, 2018.

[32] National Institute of Standards and Technology, “NIST ADDITIVE MANUFACTURING MATERIAL DATABASE,” 2019. [Online]. Available: Ammd.nist.gov.

[33] M. Yihua and X. Tuo, “Research of 4M1E’s effect on engineering quality based on structural equation model,” *Syst. Eng. Procedia*, vol. 1, pp. 213–220, 2011.

[34] H. Ko, P. Witherell, Y. Ndiaye, and Y. Lu, “Machine Learning based Continuous Knowledge Engineering for Additive Manufacturing,” in *2019 IEEE 15th International Conference on Automation Science and Engineering*, 2019.

Precise evaluation of GaAs/AlGaAs 129 kΩ and 1 MΩ quantum Hall array devices for a quantum Wheatstone bridge

T. Oe[1, 2], A. R. Panna[2], R. E. Elmquist[2], D. G. Jarrett[2], Y. Fukuyama[1], and N.-H. Kaneko[1]
[1]National Metrology Institute of Japan (NMIJ), Advanced Industrial Science and Technology (AIST), Tsukuba, 305-8563, Japan
t.oe@aist.go.jp
[2]National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, 20899, USA
alireza.panna@nist.gov

***Abstract*—We have fabricated 129 kΩ and 1 MΩ quantum Hall array devices using GaAs/AlGaAs heterostructures and their quantized Hall resistance plateaus were precisely evaluated with an accuracy of better than 5×10^{-8} based on NIST's quantized Hall resistance standard via stable standard resistors. These higher quantized values have been attracting industrial interests for applications such as precise small electric power (small current) measurements, leakage current, and electrical insulation measurements. Given these quantum Hall array devices, an accuracy of better than 1×10^{-7} for 10 MΩ resistance standard measurements will be achieved, through a simple cryogenic quantum Wheatstone bridge technique. These devices could also be utilized for a quantum voltage divider.**

***Index Terms*—quantum Hall effect, quantum Hall array resistance standard (QHARS), high resistance measurement, low current measurement, GaAs/AlGaAs heterostructure, quantum Wheatstone bridge.**

I. Introduction

Precise measurement of DC high resistance is important for the evaluation of insulation materials, measurements of small leakage currents for the development of low power consumption devices, current-based measurements of ionizing radiation, etc. High resistance values have been measured precisely using a two-terminal cryogenic current comparator (CCC) bridge [1], [2], a four-terminal CCC [3], or an active arm bridge [4], [5] in primary metrology laboratories. A quantized Wheatstone bridge, consisting of quantum Hall array devices, could provide another way to measure 10 MΩ resistance values precisely and to confirm the reliability of these other existing measurement methods. Quantum Wheatstone bridges have been used for universality comparisons of the quantum Hall effect [6] and have been proposed for precise four-terminal measurements [7] but these treated only 1 : 1 ratios and there have been no reports regarding a 10 : 1 quantum Wheatstone bridge to date. In this paper, we propose a 10 : 1 quantum Wheatstone bridge for the precise measurement of 10 MΩ with relative uncertainty better than 1×10^{-7}. The bridge consists of a 12.9 kΩ single Hall device, a 129 kΩ device which contains ten Hall bar elements connected in series, and a 1 MΩ quantum Hall array device, and we report the precise measurement results of the 129 kΩ and the 1 MΩ devices.

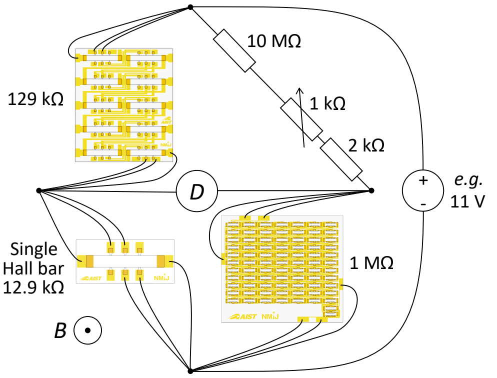


Fig. 1. Conceptual diagram of the quantum Wheatstone bridge. The left side realizes a precise and almost exact 10 : 1 ratio using the 10 series quantum Hall array device and the single Hall device. All quantum Hall array devices are connected to each other using the triple series connection technique to minimize the influence of the wiring and contact resistance.

II. Quantum Wheatstone bridge and the precise evaluation of array devices

Figure 1 shows the conceptual diagram of the quantum Wheatstone bridge. The bridge consists of a 12.9 kΩ single Hall device, a 129 kΩ device which contains ten Hall bar elements connected in series to realize a precise and almost exact 10 : 1 ratio in the left arm, and has a 1 MΩ quantum Hall array device in the right arm. A 10 MΩ standard resistor is connected in the upper right with a serially connected 2 kΩ standard resistor and a 1 kΩ precise variable decade resistor box with the minimal step of 100 mΩ to balance the bridge and to realize 10^{-8} resolution. The 10 MΩ resistance value can be determined precisely by adjusting the resistance value of the variable resistor and measuring the residual voltage or current between the neutral points. In the diagram, all array devices are connected to each other using the triple series connection technique to minimize the influence of wiring and contact resistances.

Figure 2 shows the photographs of 129 kΩ and 1 MΩ array devices mounted on TO-8 chip carriers. The 129 kΩ array device consists of ten series-connected Hall bars. The 1 MΩ array device consists of 79 Hall bars, with 77 Hall bars connected in series and two Hall bars connected in parallel

Oe, Takehiko; Panna, Alireza; Elmquist, Randolph; Jarrett, Dean G.; Fukuyama, Yasuhiro; Kaneko, Nobu-Hisa. "Precise evaluation of the 129 kΩ and 1 MΩ quantum Hall array devices for a quantum Wheatstone bridge." Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

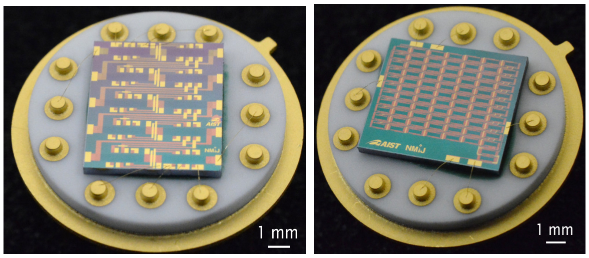


Fig. 2. Photograph of the (a) 129 kΩ and (b) 1 MΩ quantum Hall array devices mounted on a TO-8 chip carriers. 10 and 79 Hall bars are integrated on the chips, respectively.

to realize the ratio of 77.5 times $R_K/2$ and its quantized resistance value is $h/2e^2 \times 77.5 \simeq 1000\ 246.289\ \Omega$.

The current dependence curves of the Hall and the longitudinal resistance against the magnetic flux density and the precise measurement results using a four terminal cryogenic current comparator (CCC) bridge of the 129 kΩ and 1 MΩ quantum Hall array device were obtained. The results for the 1 MΩ device are shown in Figures 3 and 4. The devices were cooled down using a ^{3}He refrigerator and the device temperature was determined from the vapor pressure of ^{3}He gas to be less than 0.5 K. Both devices show flat and wide $\nu = 2$ plateaus and their longitudinal resistance shows no noticeable change for applied source-drain voltages from 1 V to 10 V. From these results, it is expected that the 10 : 1 quantum Wheatstone bridge shown in Fig. 1 can operate with an applied voltage of 11 V (77.5 μA for the left side). Both devices were fabricated on the same 20 mm square chip so that they have similar carrier density, and these devices showed no noticeable magnetic dependence from 9.3 T to 10.3 T for the 129 kΩ and from 8.4 T to 9.9 T for the 1 MΩ device. The results show that these devices can be used at the same magnetic flux density in the same refrigerator.

Since the 129 kΩ and the 1 MΩ array devices were quantized to well within 10 nΩ/Ω of their designed values, it is expected that a 10 MΩ standard resistor can be evaluated or measured precisely with relative uncertainty less than 1×10^{-7} by using the quantum Wheatstone bridge. If the applied voltage to the bridge is 11 V and the input impedance of the current detector is 10^4 Ω with a transimpedance gain of 10^{11} V/A, a resolution of 1×10^{-8} can be achieved by detecting 0.09 V (0.9 pA) at the output of the detector. The insulation resistance of the wiring should be over 10^{15} Ω for the precise measurement of 10 MΩ standards. To ensure this insulation resistance, an active guard might be needed.

III. Conclusion

The 129 kΩ and 1 MΩ GaAs/AlGaAs quantum Hall array devices were fabricated and evaluated precisely using the cryogenic current comparator bridge. These devices were well quantized to within 10 nΩ/Ω of their designed value for an overlapping range of magnetic field from 9.3 T to 9.9 T. The quantum Wheatstone bridge measurement with these devices will be performed.

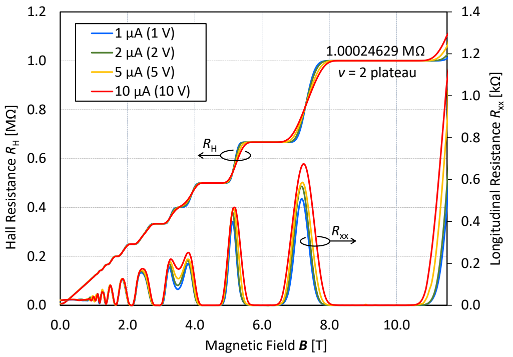


Fig. 3. Current dependence of the Hall and longitudinal resistance for the 1 MΩ quantum Hall array device (V96) from 1 μA (1 V) to 10 μA (10 V).

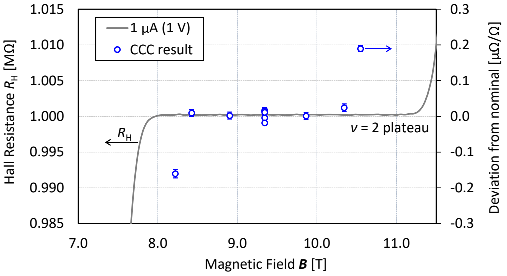


Fig. 4. Magnetic field dependence of the 1 MΩ quantum Hall array device (V96). The gray solid line shows the Hall resistance curve using a DVM sweeping the magnetic field. The blue circles shows the precise measurement results using the CCC with an applied current of 1 μA. The error bar shows the Type A uncertainty.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP18H05258 and JP18H01885. The quantum Hall array devices were fabricated at the AIST Nano-Processing Facility, supported by "Nanotechnology Platform Program" of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

[1] F. L. Hernandez-Marquez, M. E. Bierzychudek, G. R. Jones Jr., and R. E. Elmquist, "Precision high-value resistance scaling with a two-terminal cryogenic current comparator," *Rev. Sci. Instrum.*, vol. 85, 044701, April 2014.

[2] R. E. Elmquist, E. Hourdakis, D. G. Jarrett, and N. M. Zimmerman, "Direct resistance comparisons from the QHR to 100 MΩ using a cryogenic current comparator," *IEEE Trans. Instrum. Meas.*, vol. 54, pp. 525-528, April 2005.

[3] S. P. Giblin, "Re-evaluation of uncertainty for calibration of 100 MΩ and 1 GΩ resistors at NPL," *arXiv:1808.09214v1*, August 2018.

[4] L. Henderson, "A new technique for the automatic measurement of high value resistors," *J. Phys. E: Sci. Instrum.*, vol. 20, pp. 492-495, September 1987.

[5] D. G. Jarrett, "Analysis of a dual-balance high-resistance bridge at 10 TΩ," *IEEE Trans. Instrum. Meas.*, vol. 50, pp. 249-254, April 2001.

[6] F. Schopfer and W. Poirier, "Testing universality of the quantum Hall effect by means of the Wheatstone bridge," *J. Appl. Phys.*, vol. 102, 054903, July 2007.

[7] M. Marzano, L. Callegaro, and M. Ortolano, "A quantum Hall effect Kelvin bridge for resistance calibration," *IEEE CPEM Conf. Dig.*, July 2018.

Phononic Frequency Combs For Engineering MEMS/NEMS Devices With Tunable Sensitivity

Adarsh Ganesan[1*], Ashwin Seshia[2], and Jason J. Gorman[1]
[1]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]The Nanoscience Centre, University of Cambridge, Cambridge CB3 0FF, UK
*adarsh.ganesan@nist.gov

***Abstract*—Over the past two decades, MEMS resonators have received considerable attention for physical, chemical and biological sensing applications. Typically, the operation of MEMS resonant sensors relies on the tracking of a resonance frequency using a feedback oscillator. The sensitivity of these sensors is limited by physical parametric variations, as in the Young's modulus, and noise in the oscillator circuit, such that improvement in the sensitivity can require significant effort in the design, fabrication, ovenization, and control of the resonator. In this paper, we experimentally demonstrate an alternative sensing approach based on a newly documented physical phenomenon, 'phononic frequency combs', where the sensitivity can be actively tuned by the drive conditions. In addition, the spectral response of frequency combs enables an 'N+1' fold enhancement in the sensitivity, with '2N+1' being the number of spectral lines associated with a frequency comb.**

Keywords—MEMS resonator; mechanical resonator; sensor; temperature sensor; nonlinear dynamics; phononic frequency comb.

I. Introduction

MEMS resonators have been actively explored for sensing applications by many research groups around the world [1-8]. Resonant sensing is traditionally made possible by tracking the resonance frequency of one of the modes of the MEMS structure with the help of an electronic feedback circuit. Parametric variations due to temperature fluctuations and noise in the oscillator circuit can limit the performance of these sensors. Solutions for mitigating these limits, including the use of material combinations that minimize parametric variations, ovenization, and improved circuit and control design are possible but require significant research effort to optimize their performance. In this paper, we present an alternative sensing approach that fundamentally relies on the nonlinear dynamics of MEMS resonators leading to the generation of phononic frequency combs. Phononic frequency combs correspond to a series of equidistant and phase-coherent frequencies [9-18]. These frequency combs can be generated when a resonant

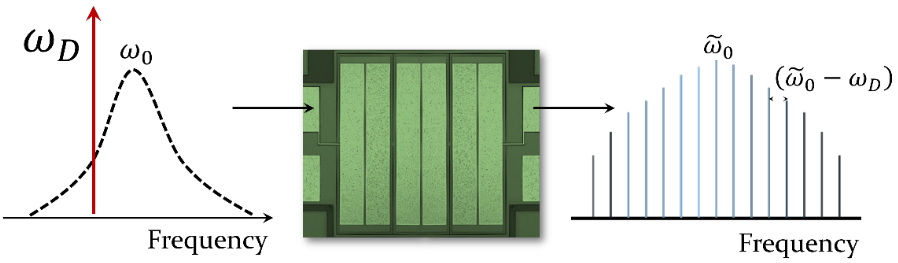


Figure 1: A strong off-resonance drive of the mechanical mode leads to an array of equidistant frequencies, referred to as a 'phononic frequency comb'. The equidistant frequency spacing of the frequency comb is set by the difference between drive frequency (ω_D) and the renormalized resonant frequency ($\widetilde{\omega}_0$) of the MEMS resonator.

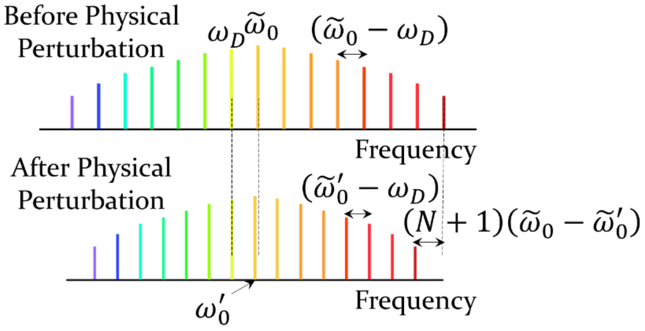

Before Physical Perturbation
$\omega_D \widetilde{\omega}_0$
$(\widetilde{\omega}_0 - \omega_D)$
Frequency
After Physical Perturbation
$(\widetilde{\omega}'_0 - \omega_D)$
$(N+1)(\widetilde{\omega}_0 - \widetilde{\omega}'_0)$
ω'_0
Frequency

Figure 2: The shift of the frequency comb line $[\widetilde{\omega}_0 + N(\widetilde{\omega}_0 - \omega_D)]$ is 'N+1' times more sensitive than the fundamental comb line $[\widetilde{\omega}_0]$.

Ganesan, Adarsh; Seshia, Ashwin; Gorman, Jason J. "Phononic Frequency Combs For Engineering MEMS/NEMS Devices With Tunable Sensitivity." Paper presented at IEEE SENSORS 2019, Montreal, CA. October 27, 2019 - October 30, 2019.

mode of a MEMS resonator is strongly driven off-resonance (Figure 1). The frequency spacing of such combs is governed by the difference between the drive frequency (ω_D) and the renormalized resonance frequency ($\widetilde{\omega}_0$) of the MEMS resonator (Figure 1). Unlike the resonance frequency of MEMS resonators (ω_0), the renormalized resonance frequency can heavily depend on the drive conditions. We want to leverage this dependence for sensing applications. Also, since the shift of the renormalized resonance frequency is cumulatively accrued in the highest order frequency comb lines, an 'N+1' fold enhancement in the sensitivity can be obtained: $[\widetilde{\omega}_0 + \mathrm{N}(\widetilde{\omega}_0 - \omega_D)] - [\widetilde{\omega}'_0 + \mathrm{N}(\widetilde{\omega}'_0 - \omega_D)] = \left[\underbrace{(\mathrm{N}+1)} * (\widetilde{\omega}_0 - \widetilde{\omega}'_0)\right]$, as pictorially demonstrated in Figure 2.

II. Experimental Setup

For the demonstration of the phononic frequency comb-based sensing approach, a MEMS system containing three coupled free-free beam structures of dimensions $1100\ \mu\mathrm{m} \times 350\ \mu\mathrm{m} \times 11\ \mu\mathrm{m}$ is considered (Figure 1). The mechanical coupler connecting the three beams has the dimensions $20\ \mu\mathrm{m} \times 2\ \mu\mathrm{m} \times 11\ \mu\mathrm{m}$. The device consists of three material layers: $10\ \mu\mathrm{m}$ thick Si, $0.5\ \mu\mathrm{m}$ thick AlN and $1\ \mu\mathrm{m}$ thick Al. The Al pad electrodes that are connected to Al patterns on the MEMS device allow for electrical interfacing, and the AlN film provides piezoelectric actuation in response to the electrical signal inputs. The synthesized electrical signal from an arbitrary waveform generator with the frequency set close to the resonance frequency is applied and the resulting response of the device is characterized using a spectrum analyzer. Additionally, a vector network analyzer is used to characterize the linear response of MEMS resonator. All frequency measurements have an uncertainty of +/- 8 Hz. The entire set of measurements presented here was conducted in a temperature-controlled oven.

III. Temperature Sensitivity of the Linear Dynamics for the MEMS Resonator

The open-loop response of the MEMS resonator is obtained at 0 dBm (Figure 3A). The peak frequency of ≈ 3.853 MHz is tracked using a manual peak detection algorithm at varying oven temperature and the scatter in the data is due to measurement repeatability and fit errors (Figure 3B). The temperature range of 70 ℃ to 90 ℃ was chosen arbitrarily. The temperature coefficient of frequency (TCF) for this specific device under test is calculated to be ≈ -4.67 Hz/MHz/℃ (ppm/℃).

IV. Temperature Sensitivity of Phononic Frequency Combs

For the excitation of phononic frequency combs, a single tone is fed into the MEMS resonator. The device is set at a temperature of 70 ℃ and the drive conditions corresponding to the existence of phononic combs is charted out in Figure 4. A representative spectral response of the phononic frequency comb is presented in Figure 5, with frequencies at $\widetilde{\omega}_0 + \mathrm{N}(\widetilde{\omega}_0 - \omega_D)$.

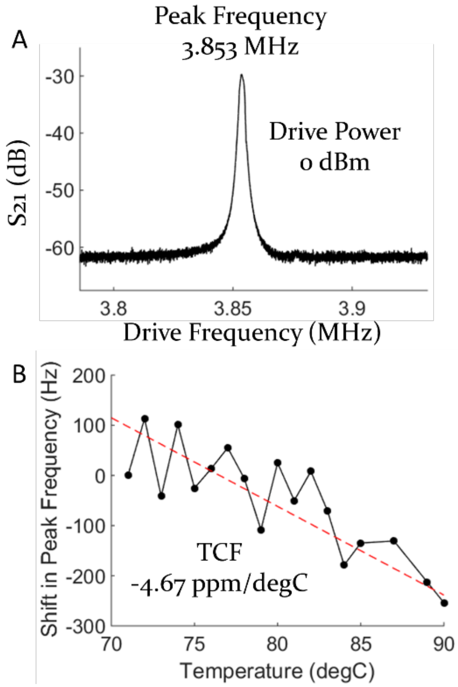


Figure 3: A: Linear response of the MEMS resonator, as recorded using the network analyzer; B: The shift in the peak frequency as the temperature is varied from 70 ℃ to 90 ℃. A linear fit to the experimental data is used to calculate the temperature coefficient of frequency (TCF).

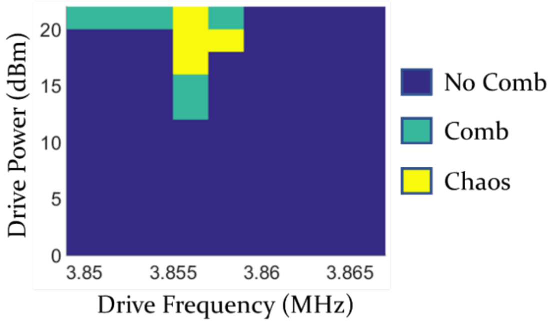


Figure 4: The drive parameter map indicating the existence of phononic frequency combs - 'green' colored region.

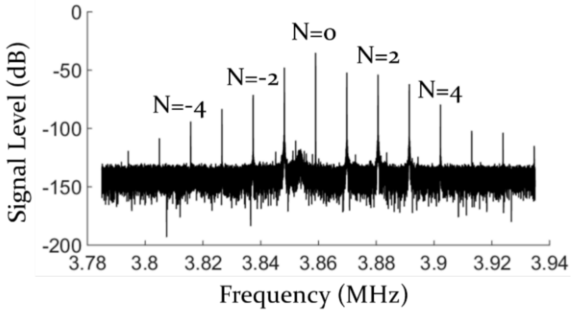


Figure 5: The spectrum of phononic frequency combs at the drive frequency 3.859 MHz and drive power 22 dBm.

Ganesan, Adarsh; Seshia, Ashwin; Gorman, Jason J. "Phononic Frequency Combs For Engineering MEMS/NEMS Devices With Tunable Sensitivity." Paper presented at IEEE SENSORS 2019, Montreal, CA. October 27, 2019 - October 30, 2019.

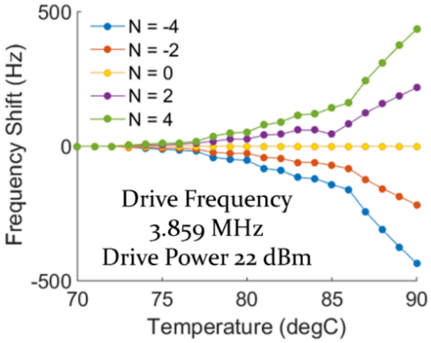


Figure 6: Temperature dependence of different lines, $\mathrm{N} = 0, \pm 2, \pm 4$, of the phononic frequency comb: $[\widetilde{\omega}_0 + \mathrm{N}(\widetilde{\omega}_0 - \omega_D)]$.

Figure 6 shows the temperature dependence of different comb lines of the phononic frequency comb. The higher order comb lines of $\mathrm{N} = \pm 4$ demonstrate greater shifts in frequency than those associated with the nearer lines $\mathrm{N} = \pm 2$. While the frequency comb in this experiment contains a sparser number of comb lines (Figure 5), it is desirable to engineer a broadband comb spectrum to achieve greater sensitivity by tracking the highest order line in the comb using narrowband electrical filtering. From Figure 6, we can see that there is a greater temperature gradient in the range of 85 ℃ to 90 ℃. Hence, for greater temperature sensitivity, it is desirable to operate the device in this band of temperatures, which can be practically implemented using *in situ* microheaters.

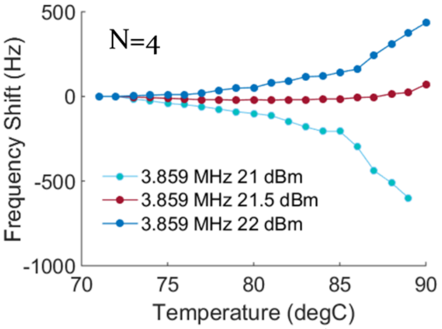


Figure 7: Temperature dependence of the 4th line of the phononic frequency comb $[\widetilde{\omega}_0 + \mathrm{N}(\widetilde{\omega}_0 - \omega_D)]$ at varying drive conditions.

While it is desirable to track the higher order comb line to achieve greater sensitivity (Figure 6), Figure 7 shows that even a slight increase in the drive power from 21 dBm to 22 dBm can reverse the nature of the temperature dependence, with the drive power of 21.5 dBm leading to temperature-insensitive frequency comb lines. Hence, without the need for sophisticated materials, fabrication, and ovenization, the temperature sensitivity can be actively tuned by the drive conditions. In comparison to Figure 3B, the frequency shift over the temperature range of 70 ℃ to 90 ℃ has increased by a factor of 5/3, which in turn can be improved by rigorous computational and experimental studies of the dynamics of phononic frequency combs.

V. SUMMARY

This paper demonstrates that the temperature sensitivity of MEMS resonators can be tuned using a newly documented physical mechanism: phononic frequency combs. The highest order line in the frequency combs offers 'N+1' fold enhancement in the sensitivity, with '$2\mathrm{N}+1$' being the number of comb lines. The sensitivity of these comb lines can be further tuned by the drive power and drive frequency. While we chose temperature as the physical variable to prove the concept of frequency comb-based sensing, the technique can be readily used for other sensing applications including particle detection. The concept of frequency comb-sensing is also relevant to other established frequency comb sources including microresonator-based frequency combs [19-22]. By integrating such multiwavelength optical sources with phononic frequency combs through dynamic optomechanical interactions, the number of comb lines $(2\mathrm{N}+1)$ can be drastically enhanced to achieve a greater magnitude of '$\mathrm{N}+1$' fold enhancement in the sensitivity.

REFERENCES

[1] Howe, R.T. and Chang, S.C., Massachusetts Institute of Technology, 1989. Resonant accelerometer. U.S. Patent 4,851,080.

[2] Yazdi, N., Ayazi, F. and Najafi, K., 1998. Micromachined inertial sensors. Proceedings of the IEEE, 86(8), pp.1640-1659.

[3] Brand, O., Dufour, I., Heinrich, S., Heinrich, S.M., Josse, F., Fedder, G.K., Korvink, J.G., Hierold, C. and Tabata, O. eds., 2015. Resonant MEMS: fundamentals, implementation, and application. John Wiley & Sons.

[4] Norouzpour-Shirazi, A., Zaman, M.F. and Ayazi, F., 2014. A digital phase demodulation technique for resonant MEMS gyroscopes. IEEE Sensors Journal, 14(9), pp.3260-3266.

[5] Seshia, A.A., Palaniapan, M., Roessig, T.A., Howe, R.T., Gooch, R.W., Schimert, T.R. and Montague, S., 2002. A vacuum packaged surface micromachined resonant accelerometer. Journal of Microelectromechanical systems, 11(6), pp.784-793.

[6] Jha, C.M., Bahl, G., Melamud, R., Chandorkar, S.A., Hopcroft, M.A., Kim, B., Agarwal, M., Salvia, J., Mehta, H. and Kenny, T.W., 2007, June. CMOS-compatible dual-resonator MEMS temperature sensor with milli-degree accuracy. In TRANSDUCERS 2007-2007 International Solid-State Sensors, Actuators and Microsystems Conference (pp. 229-232). IEEE.

[7] Lynch, J.P., Partridge, A., Law, K.H., Kenny, T.W., Kiremidjian, A.S. and Carryer, E., 2003. Design of piezoresistive MEMS-based accelerometer for integration with wireless sensing unit for structural monitoring. Journal of Aerospace Engineering, 16(3), pp.108-114.

[8] Hui, Y., Nan, T., Sun, N.X. and Rinaldi, M., 2014. High resolution magnetometer based on a high frequency magnetoelectric MEMS-CMOS oscillator. Journal of Microelectromechanical Systems, 24(1), pp.134-143.

[9] Ganesan, A., Do, C. and Seshia, A., 2017. Phononic frequency comb via intrinsic three-wave mixing. Physical review letters, 118(3), p.033903.

[10] Cao, L.S., Qi, D.X., Peng, R.W., Wang, M. and Schmelcher, P., 2014. Phononic frequency combs through nonlinear resonances. Physical review letters, 112(7), p.075505.

[11] Ganesan, A., Do, C. and Seshia, A., 2018. Excitation of coupled phononic frequency combs via two-mode parametric three-wave mixing. Physical Review B, 97(1), p.014302.

[12] Ganesan, A., 2018. Phononic frequency combs (Doctoral dissertation, University of Cambridge).

[13] Czaplewski, D.A., Chen, C., Lopez, D., Shoshani, O., Eriksson, A.M., Strachan, S. and Shaw, S.W., 2018. Bifurcation Generated Mechanical Frequency Comb. Physical review letters, 121(24), p.244302.

[14] Park, M. and Ansari, A., 2019. Formation, Evolution, and Tuning of Frequency Combs in Microelectromechanical Resonators. Journal of Microelectromechanical Systems, 28(3), pp.429-431.

[15] Houri, S., Hatanaka, D., Blanter, Y.M. and Yamaguchi, H., 2019. Modal Analysis Investigation of Mechanical Kerr Frequency Combs. arXiv preprint arXiv:1902.10289.

[16] Guerrieri, A., Frangi, A. and Falorni, L., 2018. An Investigation on the Effects of Contact in MEMS Oscillators. Journal of Microelectromechanical Systems, 27(6), pp.963-972.

[17] Ganesan, A. and Seshia, A., 2019. Resonant frequency tracking in a micromechanical device using phononic frequency combs. Scientific Reports.

[18] Dykman, M.I., Rastelli, G., Roukes, M.L. and Weig, E.M., 2019. Resonantly induced friction and frequency combs in driven nanomechanical systems. Physical review letters, 122(25), p.25430

[19] Kippenberg, T.J., Holzwarth, R. and Diddams, S.A., 2011. Microresonator-based optical frequency combs. science, 332(6029), pp.555-559.

[20] Del'Haye, P., Schliesser, A., Arcizet, O., Wilken, T., Holzwarth, R. and Kippenberg, T.J., 2007. Optical frequency comb generation from a monolithic microresonator. Nature, 450(7173), p.1214.

[21] Ferdous, F., Miao, H., Leaird, D.E., Srinivasan, K., Wang, J., Chen, L., Varghese, L.T. and Weiner, A.M., 2011. Spectral line-by-line pulse shaping of on-chip microresonator frequency combs. Nature Photonics, 5(12), p.770.

[22] Yao, B., Huang, S.W., Liu, Y., Vinod, A.K., Choi, C., Hoff, M., Li, Y., Yu, M., Feng, Z., Kwong, D.L. and Huang, Y., 2018. Gate-tunable frequency combs in graphene–nitride microresonators. Nature, 558(7710), p.410.

A Unified Analytical Approach to Multi-Cell LBT-Based Spectrum Sharing Systems

Yao Ma, Susanna Mosleh, and Jason Coder
CTL, National Institute of Standards and Technology
325 Broadway, Boulder, Colorado, USA

***Abstract*— Future unlicensed spectrum sharing scenarios involve multi-cell, multi-tier access of incumbent and emerging wireless systems, such as wireless local area network (WLAN), New-Radio unlicensed (NR-U), and Long-Term Evolution (LTE) with Licensed-Assisted Access (LAA). Listen-before-talk (LBT) medium access control (MAC) is a common technique used for channel sensing and access control. Despite intense research efforts, the majority of available results have not accurately analyzed multi-cell LBT with imperfect spectrum sensing. Furthermore, while past studies involved two major spectrum sharing strategies – shared cell access (SCA) and exclusive cell access (ECA), they missed a systematic comparison between the two. In this paper, we develop a unified analytical approach which maps the effects of imperfect spectrum sensing and multi-cell, multi-tier LBT to the key performance indicators (KPIs) of LAA and WLAN cells. We provide an analytical comparison between the SCA and ECA schemes, and show that the SCA can provide a significantly higher system throughput than the ECA as a function of sensing thresholds. We program the SCA and ECA algorithms with a new simulation method and implement Monte Carlo simulations, which verify our analytical results. Numerical results provide insightful observations on effects of various parameters. These results provide powerful analytical and simulation tools to evaluate the performance of multi-tier LBT coexistence systems with imperfect sensing, and effectively support coexistence system optimization.**

Index Terms: Imperfect spectrum sensing, LTE-LAA, Multi-tier coexistence, Multi-cell LBT, WLAN.

I. Introduction

Spectrum sharing in heterogeneous 4G and 5G wireless systems [1]–[4] involves multiple radio access technologies (RATs) of emerging wireless systems such as cellular Long-Term Evolution (LTE) or New Radio (NR) systems with unlicensed or License Assisted Access (LAA) [1]–[3], and incumbent services such as wireless local area networks (WLANs) [5], [6]. In these systems, various listen-before-talk (LBT) medium access control (MAC) schemes are used for channel sensing and access control.

Regarding LBT-based multi-cell spectrum sharing, there are two popular assumptions or schemes which we call shared cell access (SCA) and exclusive cell access (ECA), respectively. Both schemes use clear channel assessment (CCA) and detect the channel busy/idle states of adjacent cells. Without loss of generality, we assume that each LAA small-cell base station (SBS) or WLAN access point (AP) does the sensing, and schedules downlink transmissions in each cell (or contention zone).

In the SCA model, each SBS/AP assumes imperfect sensing and can schedule transmissions when the received intercell interference (ICI) from adjacent cells is below the sensing threshold. This model allows several cells (or links) to transmit simultaneously when the mutual ICIs are below their sensing thresholds. The SCA methods for multi-cell networks have been considered in [7]–[10], to name a few.

The ECA scheme assumes that at any time a successful transmission happens when only one link is transmitting. It is consistent with the carrier sense multiple access with collision avoidance (CSMA/CA) performance modeling provided in [11]. The ECA scheme has been assumed in [12]–[16] for spectrum sharing. Certainly, the ECA scheme applies only to a small area of highly overlapped cells, when the SBS/AP sensing thresholds are set to be adequately low. The ECA scheme can be regarded as a special case of the SCA scheme.

In the studies mentioned above, the effect of ICI sensing errors on the network throughput has not been explicitly addressed. Furthermore, the LBT MAC features for the SCA have not been adequately modeled or studied. For example, [7], [8], [12] did not model an important MAC metric – average channel busy time to support a successful transmission. The effect of inter-SBS/AP ICI sensing errors was only implicitly studied in [12], and the LBT MAC feature (such as multistage backoff) was not adequately modeled. We have provided a novel analytical approach to evaluate the effect of CCA errors on the LAA and WLAN network throughput in [17] assuming an ECA scheme. The result in [17] cannot be readily applied to the SCA scheme.

In summary, analyzing the effect of CCA errors on the key performance indicators (KPIs) of multi-cell LBT coexistence systems has not been satisfactorily addressed in the literature. Furthermore, there is a lack of a systematic comparison of SCA and ECA models in a unified framework. It is interesting to investigate if the SCA can provide higher KPIs than the ECA even when cells are densely overlapped. Yet we could not find such a comparative analysis available in the literature. This comparative analysis is useful, for example, on the selection of sensing threshold and MAC parameters, and the design of better spectrum sharing algorithms.

The related computer simulation work is non-trivial, because the traditional discrete-event driven simulation methods, such as those targeted for the ECA (in [11], [15]) and/or perfect channel sensing, cannot be directly used. The imperfect sensing in a multi-cell network can cause multiple nodes to implement backoff or transmission simultaneously. The sim-

Ma, Yao; Mosleh, Susanna; Coder, Jason. "A Unified Analytical Approach to Multi-Cell LBT-Based Spectrum Sharing Systems." Paper presented at the 2020 IEEE 91st Vehicular Technology Conference: VTC2020-Spring, Antwerp, BE. May 25, 2020 - May 28, 2020.

ulation shall model the resulting mis-detection and collision events in a fine time step-size of a CCA duration, and follow the multicell LBT procedure. We are not aware of an open-source or commercial software that can readily simulate such a scenario.

In this paper, we address the challenging problem of modeling and analysis for multi-cell LAA and WLAN coexistence networks with imperfect sensing of ICIs, and treat the SCA and ECA schemes in a unified framework. We highlight the novel contributions as follows:

- We develop a new method on integrating effects of spectrum sensing detection probability and multi-cell LBT MAC features, and mapping them to the system KPIs, including the channel access probability (CAP), successful transmission probability (STP), and throughput.
- We provide a unified analysis and systematic comparison of the SCA and ECA schemes for LAA and WLAN coexistence networks.
- By developing a new simulation method, we program the SCA and ECA algorithms with multi-cell and multi-tier LBT, and implement Monte Carlo simulations, which validate our analytical results.

Numerical results show the effects of critical MAC and physical layer parameters such as sensing threshold and detection probability, and demonstrate the performance improvement of the SCA scheme over the ECA scheme. This technique provides significant progress on performance modeling and analysis of multi-cell LBT coexistence schemes, and support 4G and 5G multi-cell optimization in license-assisted or unlicensed spectrum bands.

II. System Model

We consider a two-tier channel access system on a 5 GHz industrial, scientific, and medical (ISM) radio band, where there is an LTE-LAA system with C_L small cells and a WLAN system with C_W cells (aka. contention zones), which are assumed to be the emerging and incumbent services, respectively. All cells share a single wide-band channel, and use CSMA/CA type of MAC channel access procedures specified in [1], [2], [5]. Each LAA and WLAN cell is controlled by an SBS and AP, respectively, or called scheduler in general, using LBT or distributed coordination function (DCF) protocols to sense the channel activity of other cells and schedule downlink transmissions. The system model is presented in Fig. 1, where each LTE-LAA SBS or WLAN AP senses channels and schedules downlink transmission to its associated users. Each user receives downlink ICIs of neighboring SBSs and APs, shown in red lines.

Let all cells be indexed by $c \in \{1, \ldots, C_{\text{tot}}\}$, where $C_{\text{tot}} = C_L + C_W$, $c \in \{1, \ldots, C_L\}$ refers to an LAA cell, and $c \in \{C_L+1, \ldots, C_{\text{tot}}\}$ refers to a WLAN cell. There are N_c users in cell c, and pair (c, n) denotes user node n in cell c. The downlink transmit power at cell c is denoted as $P_{T,c}$.

Throughout this paper, we use subscripts L, W, I, S, F, and P to denote LAA, WLAN, idle, successful transmission, failed transmission, and payload, respectively. For cell c, we define δ, $T_{P,c}$, $T_{S,c}$ and $T_{F,c}$ as durations of idle slot, payload, successful transmission, and failed transmission, respectively; and define τ_c, $P_{I,c}$, $P_{S,c}$ and $P_{F,c}$ as the probabilities of channel access, channel idle, successful transmission, and failed transmission.

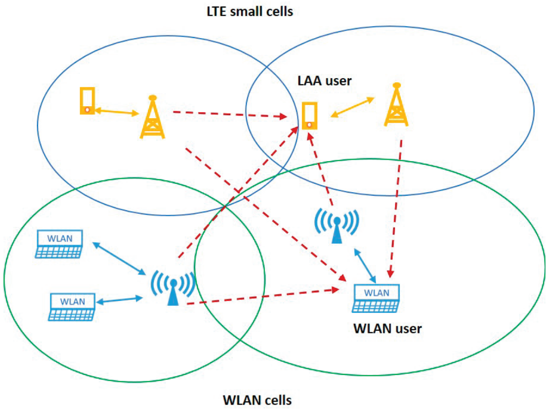


Fig. 1: Model of multi-cell LTE-LAA and WLAN systems.

We consider a frequency-flat Rayleigh block fading channel model for each communication and interference link, which is static in a block of milli-seconds level (which approximately corresponds to a transmission duration), then changes independently from block to block. The power of channel gain from scheduler c_2 to user (c, n) is denoted as $h_{(c,n),c_2}$, and the power of channel from scheduler c_2 to scheduler c is represented as g_{c,c_2}. The probability density function (PDF) of channel power $h_{(c,n),c_2}$ is given by

$$h_{(c,n),c_2}(P) = \exp(-P/\bar{P})/\bar{P}, \tag{1}$$

where $\bar{P}$ is the average power of fading channel gain. Based on [18], we have $\bar{P} = E[h_{(c,n),c_2}] = (4\pi/\lambda_c)^2 d_{(c,n),c_2}^{-\alpha_d}$, where λ_c is the carrier wavelength, $d_{(c,n),c_2}$ is distance between scheduler c_2 to node (c, n), and α_d is the path loss exponent.

Regarding the MAC parameters at scheduler c, we assume that the contention window size (CWS) is $W_{c,m}$ at backoff stage m (for $m = 0, 1, \ldots, M_c$), where M_c is the maximum backoff stage. We present the SCA and ECA schemes next.

Multi-cell Downlink SCA Scheme:

Each scheduler c implements the following procedure:

1) Based on a CSMA/CA protocol, scheduler c draws an initial backoff counter value within $(0, W_{c,m} - 1)$ based on the current CWS $W_{c,m}$.
2) Scheduler c senses channel activities and compares the received sum ICI power with sensing threshold $I_{d,c}$ during each sensing slot δ.
3) If ICI power exceeds threshold, scheduler c freezes its backoff counter; otherwise, it reduces counter value by one. Go back to Step 2) when counter is larger than zero; otherwise go to step 4).
4) Based on either request-to-send/clear to send (RTS/CTS) or basic access, scheduler c sends handshaking or payload signals to associated users.
5) If the majority of transmissions are successful, scheduler c reduces CWS to $W_{c,0}$; otherwise, it doubles the CWS,

unless the maximum CWS W_{c,M_c} is reached. Go back to step 1).

On the other hand, when the cells are highly overlapped spatially, and/or when the sensing threshold at every scheduler is very low, each scheduler can detect transmissions in any adjacent cell perfectly. We call this the Multi-cell Downlink ECA Scheduling Scheme.

To implement LBT or DCF, an SBS uses only one threshold based on energy detection (ED) $I_{L,\text{ed}}$ to sense transmissions in other cells, and an AP uses two thresholds – carrier sensing (CS) threshold $I_{W,\text{cs}}$ to sense WLAN transmissions of other cells, and ED threshold $I_{W,\text{ed}}$ to detect other type of transmissions (such as LAA signals), respectively. We use I_{d,c,c_2} to denote the sensing threshold of scheduler c to detect signals from a neighboring cell c_2, then

$$I_{d,c,c_2} = \begin{cases} I_{L,\text{ed}} & \text{for } c \in \{1, \dots, C_L\}, \text{ and any } c_2, \\ I_{W,\text{ed}} & \text{for } c \in \{C_L + 1, \dots, C_{\text{tot}}\}, \\ & \text{and } c_2 \in \{1, \dots, C_L\} \\ I_{W,\text{cs}} & \text{for } c \in \{C_L + 1, \dots, C_{\text{tot}}\}, \\ & \text{and } c_2 \in \{C_L + 1, \dots, C_{\text{tot}}\}. \end{cases}$$

Furthermore, we define $I_{f,c}$ as the ICI maximum tolerance threshold at scheduler c.

III. Performance Analysis

We first derive the multi-cell LAA and WLAN coexistence performance with the SCA, and then simplify the results to the ECA and discuss their relationship.

A. Shared Cell Access

In this method, we model the ICI from each source individually, and then combine the impacts. We use $\hat{H}_0$ and $\hat{H}_1$ to denote decisions of null and transmission, respectively. We define P_{d,c,c_2} as the detection probability of scheduler c_2 at scheduler c, when a transmission from cell c_2 with power P_{T,c_2} starts, with decision threshold I_{d,c,c_2} and local noise power spectrum density (PSD) $N_{0,c}$. Thus,

$$P_{d,c,c_2} \quad = \quad \Pr(P_{T,c_2} g_{c,c_2} + N_{0,c} < I_{d,c,c_2}).$$

Suppose that we use a sensing duration of δ_{SS} with sampling rate B_W, and obtain $[\delta_{SS} B_W]$ samples to make a block decision, where $[x]$ rounds x to its nearest integer. By default, $\delta = 9$ μs and $\delta_{SS} \geq 4$ μs at the 5 GHz ISM band [2], [5]. Typically, $\delta_{SS} B_W \gg 1$ holds. For example, when $\delta_{SS} = 5$ μs and $B_W = 20$ MHz, we have $[\delta_{SS} B_W] = 100$. With $\delta_{SS} B_W \gg 1$, the sum power of the sampled complex noise has a Gamma distribution with degree of freedom (DOF) $[2\delta_{SS} B_W]$, which approximates a constant. On the other hand, due to limited Doppler shift for small-cell communication, we model the inter-SBS/AP channel g_{c,c_2} as a slow Rayleigh fading channel, and its DOF is only 2 in the δ_{SS} sensing duration. In summary, the magnitude of the received ICI follows a Rayleigh distribution in the δ_{SS} duration (slow fading), and its power follows an exponential distribution.

We obtain the inter-cell detection probability as

$$\begin{aligned} P_{d,c,c_2} &= \Pr(P_{T,c_2} g_{c,c_2} < \tilde{I}_{d,c,c_2}) \\ &\simeq 1 - \exp(-\tilde{I}_{d,c,c_2} / [P_{T,c_2} \bar{g}_{c,c_2}]), \end{aligned} \qquad (2)$$

where $\tilde{I}_{d,c,c_2} = \max(I_{d,c,c_2} - N_{0,c}, 0)$, $\bar{g}_{c,c_2} = E[g_{c,c_2}]$, and $E[\cdot]$ here is with respect to the fading channel distribution.

Furthermore, we define P_{f,c,c_2} as the probability that ICI from cell c_2 causes transmission of cell c to fail. We obtain that $P_{f,c,c_2} = \Pr(P_{T,c_2} g_{c,c_2} < \tilde{I}_{f,c}) \simeq 1 - \exp(-\tilde{I}_{f,c} / [P_{T,c_2} \bar{g}_{c,c_2}])$, where $\tilde{I}_{f,c} = \max(I_{f,c} - N_{0,c}, 0)$.

Define $\tilde{P}_{\text{suc},c}$ as the conditional STP of scheduler c when its transmission starts, and $P_{\text{suc},c}$ as the average STP. We obtain $\tilde{P}_{\text{suc},c} = \prod_{\substack{c_2=1 \\ c_2 \neq c}}^{C_{\text{tot}}} (1 - \tau_{c_2} P_{f,c,c_2})$. Note that in $\tilde{P}_{\text{suc},c}$ the term $1 - \tau_{c_2} P_{f,c,c_2}$ can be decomposed to two parts: $1 - \tau_{c_2}$ which is probability that cell c_2 is not active and $\tau_{c_2}(1 - P_{f,c,c_2})$, which is probability that cell c_2 is active but its generated ICI does not cause transmission of cell c to fail.

The CAP of scheduler c is a function of $\tilde{P}_{\text{suc},c}, W_{c,0}$, and M_c. Based on [15], the CAP of scheduler c is given by:

$$\tau_c \quad = \quad \frac{2(1 - (1 - \tilde{P}_{\text{suc},c})^{M_c + 1})}{\tilde{P}_{\text{suc},c} \sum_{j=0}^{M_c} (1 - \tilde{P}_{\text{suc},c})^j (1 + W_c)}. \qquad (3)$$

We derive the average STP at cell c as

$$P_{\text{suc},c} = \tau_c \tilde{P}_{\text{suc},c}. \qquad (4)$$

We define $T_{\text{suc},c,n}$ as the normalized throughput duration (NTD) of user (c, n), which is given by

$$T_{\text{suc},c,n} = \alpha_{(c,n)} P_{\text{suc},c} T_{P,c} / T_{\text{ave},c},$$

where $\alpha_{(c,n)}$ is time ratio allocated to node n (with $\sum_{n=1}^{N_c} \alpha_{(c,n)} = 1$), and $T_{\text{ave},c}$ is the average time in cell c to support one successful transmission. The overall NTD at cell c is given by

$$T_{\text{suc},c} \quad = \quad \sum_{n=1}^{N_c} T_{\text{suc},c,n}. \qquad (5)$$

The NTD is conceptually similar to the successful channel occupancy efficiency, or called normalized throughput in [15]–[17]. Note that $\sum_{c=1}^{C_{\text{tot}}} T_{\text{suc},c} > 1$ can hold for the SCA scheme, but not for the ECA scheme. The NTD does not model the physical-layer data rate.

To model both MAC- and physical-layer features, we define the average throughput of user (c, n) as

$$\begin{aligned} S_{c,n} &= \alpha_{(c,n)} P_{\text{suc},c} T_{P,c} \\ &\quad \cdot B_{(c,n)} E[\log_2(1 + \beta_{(c,n)} \gamma_{(c,n)})] / T_{\text{ave},c}, \end{aligned} \qquad (6)$$

where $\gamma_{(c,n)}$ is the signal-to-interference-and-noise ratio (SINR) at user (c, n), $E[\cdot]$ is the expectation with respect to the distribution of $\gamma_{(c,n)}$, and $\beta_{(c,n)}$ is the signal-to-noise ratio (SNR) gap function [18], given by $\beta_{(c,n)} = -1.5 / \log(5\text{BER}_{(c,n)})$, where $\text{BER}_{(c,n)}$ is the target bit error rate (BER) for user traffic. The cell c sum-throughput is given by

$$S_c = \sum_{n=1}^{N_c} S_{c,n}. \qquad (7)$$

We derive $\gamma_{(c,n)}$ as

$$\gamma_{(c,n)} = \frac{P_{T,c} h_{(c,n),c}}{I_{\text{tot},(c,n)}(\hat{H}_0) + N_{0,c}}, \qquad (8)$$

where $I_{\text{tot},(c,n)}(\hat{H}_0)$ is the average total interference power at node (c,n) under $\hat{H}_0$ decision at cell c, when the scheduler c starts a downlink transmission. It is given by

$$I_{\text{tot},(c,n)}(\hat{H}_0) = \sum_{\substack{c_2=1\\c_2\neq c}}^{C_{\text{tot}}} P_{T,c_2}\tau_{c_2}\bar{h}_{(c,n),c_2}(1-P_{d,c,c_2}). \qquad (9)$$

where $\bar{h}_{(c,n),c_2} = E[h_{(c,n),c_2}]$. In (9), we introduce the factor $(1-P_{d,c,c_2})$ to represent the probability of experienced ICI at scheduler c.

Finally, we analyze the $T_{\text{ave},c}$ which is challenging to evaluate due to the multi-cell ICIs and CCA errors. To distinguish between LAA and WLAN cells, we use $T_{F,L}$ and $T_{F,W}$ to denote durations caused by failed transmission in an LAA and WLAN cell, and $T_{F,M} = \max(T_{F,L}, T_{F,W})$. After some manipulations, we derive $T_{\text{ave},c}$ as

$$\begin{aligned}T_{\text{ave},c} = P_{I,c}\delta + P_{\text{suc},c}T_{S,c} + \sum_{\substack{c_2=1\\c_2\neq c}}^{C_{\text{tot}}} P_{d,c,c_2}P_{\text{suc},c_2}T_{S,c_2}\\ +P_{F,L,c}T_{F,L} + P_{F,W,c}T_{F,W} + P_{F,LW,c}T_{F,M}, \qquad (10)\end{aligned}$$

where the first, second and third terms are related to events of system idle, successful transmission at cell c, and cell-c sensed successful transmissions of other cells, respectively; and in the 4th to 6th terms $P_{F,L,c}$, $P_{F,W,c}$, and $P_{F,LW,c}$ are, respectively, probabilities of failed transmissions (e.g. caused by collisions) at only LAA cells, at only WLAN cells, and at both LAA and WLAN cells. In detail, $P_{I,c}$ is the system idle probability sensed by scheduler c, given by

$$P_{I,c} = \prod_{c_2=1}^{C_{\text{tot}}}(1-\tau_{c_2}P_{d,c,c_2}), \qquad (11)$$

where $P_{d,c,c} = 1$ holds. In the 3rd term of (10), the factor P_{d,c,c_2} models an important fact that schedule c only freezes its counter when it can sense a downlink transmission in cell c_2. We can combine the second and third terms to $\sum_{c_2=1}^{C_{\text{tot}}} P_{d,c,c_2}P_{\text{suc},c_2}T_{S,c_2}$ by using the fact that $P_{d,c,c} = 1$ holds. Furthermore, we can obtain that in (10),

$$\begin{aligned}P_{F,L,c} &= \left[\prod_{c_2=C_L+1}^{C_{\text{tot}}}(1-\tau_{c_2}P_{d,c,c_2})\right]\\ &\cdot\left[1-\prod_{c_2=1}^{C_L}(1-\tau_{c_2}P_{d,c,c_2}) - \sum_{c_2=1}^{C_L}P_{d,c,c_2}P_{\text{suc},c_2}\right] \qquad (12)\\ P_{F,LW,c} &= \left[1-\prod_{c_2=1}^{C_L}(1-\tau_{c_2}P_{d,c,c_2})\right]\\ &\cdot\left[1-\prod_{c_2=C_L+1}^{C_{\text{tot}}}(1-\tau_{c_2}P_{d,c,c_2})\right], \qquad (13)\end{aligned}$$

where on the right handside of (12), the first factor models the probability of failed transmission event at LAA cells, and the second factor is the probability of idle channel event at WLAN cells, when both events happen and are sensed by scheduler c. The expression of $P_{F,W,c}$ can be obtained using a similar procedure, but is omitted here due to the space limitation.

When $T_{F,L} = T_{F,W} = T_F$ holds, we can simplify (10) to

$$T_{\text{ave},c} = P_{I,c}\delta + \sum_{c_2=1}^{C_{\text{tot}}} P_{d,c,c_2}P_{\text{suc},c_2}T_{S,c_2} + P_{F,c}T_F,$$

where

$$\begin{aligned}P_{F,c} &= P_{F,L,c} + P_{F,W,c} + P_{F,LW,c}\\ &= 1 - P_{I,c} - \sum_{c_2=1}^{C_{\text{tot}}} P_{d,c,c_2}P_{\text{suc},c_2}. \qquad (14)\end{aligned}$$

The novelty of our method can be briefly explained as follows: Our method is more accurate than several state-of-the-art results on modeling LBT MAC features and the impact of inter-SBS/AP ICI sensing errors. For example, [7]–[9], [12] did not model the CSMA/CA multistage backoff and/or average time to support one successful transmission, see (10). Furthermore, the provided throughput formulas in [7]–[13] did not explicitly address imperfect ICI sensing and its impact of KPIs. More generally, from the open literature we have not found validated multi-cell multi-tier KPI results for coexistence systems (such as LAA and WLAN) that flexibly take into account both CSMA/CA LBT MAC procedures and CCA sensing errors. These issues have been addressed in this paper. We show in Subsection III-B that our SCA result simplifies to that of ECA when the inter-cell detection probability approaches unity, and is equivalent to known results. In Section IV, we further provide Monte Carlo simulation results to validate the analysis.

B. Exclusive Cell Access

The ECA can be modeled by assuming perfect sensing $P_{d,c,c_2} = 1$ for all c and c_2. Certainly, this assumption is only valid for special cases, such as a region of highly overlapped cells, high transmission powers, and/or very sensitive sensing thresholds.

The average throughput at user (c,n) is given in the form of (6), but with the following changes: the τ_c, $\tilde{P}_{\text{suc},c}$, $P_{\text{suc},c}$, $\gamma_{(c,n)}$ and $T_{\text{ave},c}$ therein shall be changed by replacing $P_{d,c,c_2} = 1$ therein. Then we obtain

$$\tilde{P}_{\text{suc},c}^{\text{ECA}} = \prod_{\substack{c_2=1\\c_2\neq c}}^{C_{\text{tot}}}(1-\tau_{c_2}),$$

$P_{\text{suc},c}^{\text{ECA}} = \tau_c\tilde{P}_{\text{suc},c}^{\text{ECA}}$ and $\gamma_{(c,n)}^{\text{ECA}} = \frac{P_{T,c}h_{(c,n),c}}{N_{0,c}}$. Note that the SINR $\gamma_{(c,n)}$ in the SCA scheme becomes the SNR in the ECA scheme since ICI is avoided due to perfect inter-cell sensing.

To gain insight into LAA and WLAN coexistence, we assume that the C_L LAA cells have homogeneous CSMA/CA parameters, and so are the C_W WLAN cells. We also write the CAP, the idle, successful transmission, and failed transmission probabilities of all the LAA (or WLAN) cells as τ_L, $P_{I,L}$, $P_{S,L}$, and $P_{F,L}$ (or τ_W, $P_{I,W}$, $P_{S,W}$, and $P_{F,W}$), respectively.

Under assumptions of perfect sensing and ECA, we obtain:

$$\begin{aligned} P_{I,L} &= \prod_{c=1}^{C_L} (1-\tau_{c,L}), \\ P_{S,L} &= \sum_{c=1}^{C_L} \tau_{c,L} \left[\prod_{\substack{c_2=1 \\ c_2 \neq c}}^{C_L} (1-\tau_{c_2,L}) \right], \\ P_{F,L} &= 1 - P_{I,L} - P_{S,L}. \end{aligned}$$

When all SBS nodes have homogenous MAC parameters, i.e., $\tau_{c,L} = \tau_L$ for all c, we obtain:

$$\begin{aligned} P_{I,L} &= (1-\tau_{c,L})^{C_L}, \\ P_{S,L} &= C_L \tau_{c,L} (1-\tau_{c,L})^{C_L - 1}, \end{aligned}$$

and $P_{F,L} = 1 - P_{I,L} - P_{S,L}$. Similarly, $P_{I,W} = (1-\tau_{c,W})^{C_W}$,$P_{S,W} = C_W \tau_{c,W} (1-\tau_{c,W})^{C_W - 1}$, and $P_{F,W} = 1 - P_{I,W} - P_{S,W}$.

Furthermore, (10) is simplified to

$$\begin{aligned} T_{\text{ave}}^{\text{ECA}} = P_I^{\text{ECA}} \delta + P_{\text{suc},L}^{\text{ECA}} T_{S,L} + P_{\text{suc},W}^{\text{ECA}} T_{S,W} \\ + P_{F,L}^{\text{ECA}} T_{F,L} + P_{F,W}^{\text{ECA}} T_{F,W} + P_{F,LW}^{\text{ECA}} T_{F,M}, \qquad (15) \end{aligned}$$

where $P_I^{\text{ECA}} = P_{I,L} P_{I,W}$, $P_{\text{suc},L}^{\text{ECA}} = P_{I,W} P_{S,L}$, $P_{\text{suc},W}^{\text{ECA}} = P_{I,L} P_{S,W}$, $P_{F,L}^{\text{ECA}} = P_{I,W} P_{F,L}$, $P_{F,W}^{\text{ECA}} = P_{I,L} P_{F,W}$, and $P_{F,LW}^{\text{ECA}} = (1-P_{I,L})(1-P_{I,W})$. We can show that (15) is equivalent to a result given by eq. (19) in [15], when the latter is reduced from three transmission types to two types.

IV. Numerical Results

In this section, we provide both analytical and simulation results to validate our analysis, and show impact of critical parameters such as the sensing thresholds.

We develop a new simulator by using a constant simulation step-size (aka, event update interval) of an idle slot duration δ to track the effect of sensing error event. All schedulers (aka. SBSs and APs) follow multi-cell LBT procedures. We assume a slow fading channel, and the sensing result at each scheduler is updated when any transmission in the system starts or stops. The simulator tracks all the backoff, transmission, and sensing error events, and computes average KPIs.

Some CSMA/CA parameters and equations to compute $T_{S,L}$ (and $T_{F,L}$) from $T_{P,L}$, and to compute $T_{S,W}$ (and $T_{F,W}$) from $T_{P,W}$ are provided in [16]. Assume that there are N_L (or N_W) users in each LAA (or WLAN) cell, with CWS $W_{L,0}$ (or $W_{W,0}$), and maximum backoff stage M_L (or M_W), respectively. Unless otherwise stated, we assume that $\alpha_d = 3.5$, $\delta = 9\mu s$, total area has a rectangle shape with size $R_0 \times R_0$, and each cell has a disk shape with radius r_0. Further, we assume carrier frequency $f_c = 5.2$ GHz, $P_{T,c} = 23$ dBm, $B_{(c,n)} = 20$ MHz, $\text{BER}_{(c,n)} = 10^{-3}$, $\alpha_{(c,n)} = 1/N_c$ for all c and n, and background white Gaussian noise has PSD of -174 dBm/Hz. RTS/CTS is used for both LAA and WLAN cells. To model the effects of minimum link distance and limited modulation size, the SNR or SINR per link is truncated to 20 dB as an upper-bound. We choose these physical and MAC parameters to represent a feasible multicell coexistence scenario.

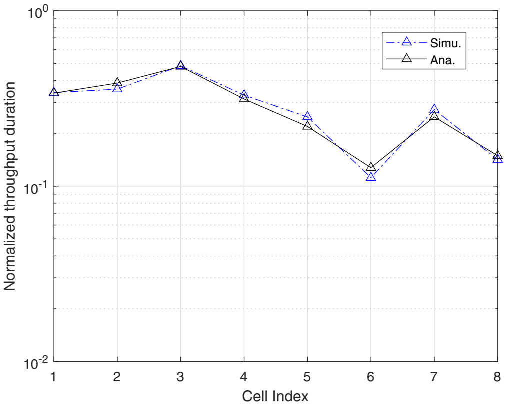


Fig. 2: Normalized cell-throughput duration of the LTE-LAA and WLAN system vs. cell index, and the first and next 4 cells are LAA and WLAN cells, respectively.

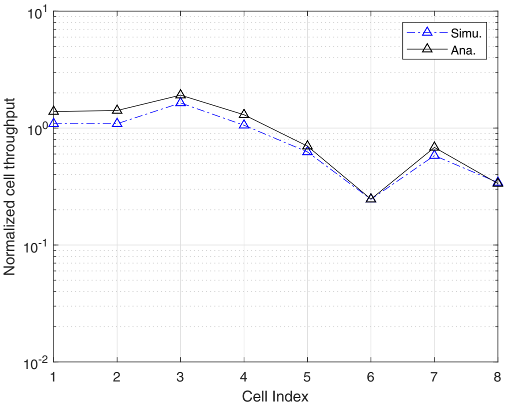


Fig. 3: Normalized cell throughput of the LTE-LAA and WLAN system vs. cell index, with the same system setting as for Fig. 2.

We provide analytical and simulation results on the per-cell NTD based on eq. (5) in Fig. 2, and throughput based on eq. (7) but normalized by the channel bandwidth in Fig. 3. For both figures, we assume that $N_L = N_W = 5$, $W_{L,0} = 8$, $W_{W,0} = 16$, $M_L = M_W = 1$, $I_{f,c} = I_{d,c} = -72$ dBm for all c, $T_{P,W} = T_{P,L} = 10\delta$, $R_0 = 100$ m, and $r_0 = 30$ m. We ran the algorithms for 2×10^4 time slots to obtain the average statistics. Due to random locations of SBSs, APs and users, the NTDs and normalized throughputs among the cells are heterogeneous. Figs. 2 and 3 illustrate consistent matching among analytical and simulation results, which account for effects of ICI and sensing errors. The minor mismatch between simulation and analytical results in Fig. 3 is likely caused by the difficulty of analytically evaluating the distribution of the sum ICIs experienced by each user when packets are received. This issue will be addressed in our future work.

Next, we show analytical numerical results on normalized

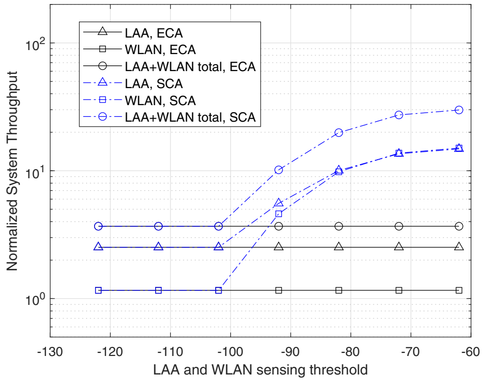


Fig. 4: Normalized system throughput of the LTE-LAA and WLAN systems vs. LAA and WLAN sensing threshold.

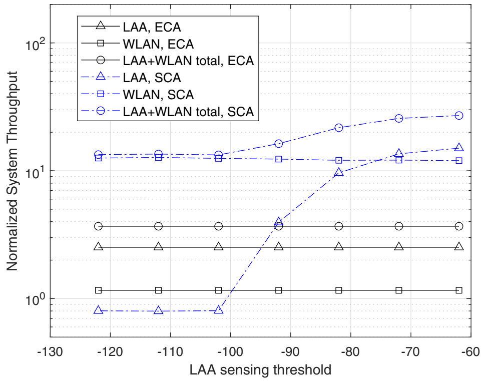


Fig. 5: Normalized system throughput of the LTE-LAA and WLAN systems vs. only LAA sensing threshold.

system throughput vs. sensing thresholds in Fig. 4 and Fig. 5, respectively, assuming $C_L = C_W = 6$, $N_L = N_W = 5$, $W_{L,0} = W_{W,0} = 16$, $M_L = 1$, $M_W = 4$, $I_{L,\text{ed}} = -72$ dBm, $I_{W,\text{cs}} = -82$ dBm, $I_{W,\text{ed}} = -62$ dBm, $T_{P,W} = T_{P,L} = 1$ ms, $R_0 = 200$ m and $r_0 = 30$ m. The results were obtained by averaging over 1000 random location profiles.

For Fig. 4, we assume that all LAA and WLAN sensing thresholds are equal and vary between [-122,-62] dBm. Fig. 4 verifies that when all sensing thresholds are reduced to about -102 dBm or lower (from the right to the left), the throughputs of the SCA scheme (LAA and WLAN) are reduced significantly and converge to those of the ECA scheme, as expected, because the inter-SBS/AP detection probabilities approach unity in this example. Fig. 5 provides throughput vs. LAA sensing threshold $I_{L,\text{ed}}$ which varies in the range of [-122,-62] dBm, but WLAN sensing thresholds are fixed. We observe that when LAA SBSs reduce their sensing threshold from -62 dBm to -102 dBm, the LAA throughput reduces significantly from about 15 to 0.8, but the WLAN throughput is insensitive to the LAA threshold change.

V. Conclusion

In this paper, we have modeled and analyzed the KPIs of LBT-related multi-cell 2-tier coexisting systems (LTE-LAA and WLAN), and provided a unified analysis of the multi-cell SCA and ECA schemes as function of detection probability and sensing thresholds. We have programmed an idle-slot step-size event-update simulation tool to better track sensing error events and provided reliable simulation results to verify our analysis. Numerical results demonstrate effects of various MAC and system parameters, such as LAA and WLAN sensing thresholds, and show that SCA with proper threshold setting provides significantly greater cell sum throughput than the ECA scheme. This result provides powerful analytical and simulation tools for performance evaluation of multi-tier coexistence systems with imperfect sensing, and supports system optimization in a practical way.

References

[1] 3GPP TSG RAN, “Study On Licensed-Assisted Access To Unlicensed Spectrum”, 3GPP TR 36.889 V13.0.0, Jun. 2015.

[2] 3GPP TS RAN, “E-UTRA Physical layer procedures (Release 14)”, 3GPP TS 36.213 V14.4.0, Sept. 2017.

[3] 3GPP TSG RAN, “Study on NR-based access to unlicensed spectrum”, 3GPP TR 38.889 V16.0.0, Dec. 2018.

[4] A. Mukherjee et al., “Licensed-assisted access LTE: coexistence with IEEE 802.11 and the evolution toward 5G,” *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 50-57, Jun. 2016.

[5] IEEE LAN/MAN Standards Committee, IEEE Std 802.11-2012, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Feb. 2012.

[6] L. Dai and X. Sun, “A unified analysis of IEEE 802.11 DCF networks: stability, throughput, and delay,” *IEEE Trans. Mobile Computing*, vol.12, no.8, pp.1558–1572, Aug. 2013.

[7] Y. Li, F. Baccelli, J. G. Andrews, T. D. Novlan and J. C. Zhang, “Modeling and analyzing the coexistence of Wi-Fi and LTE in unlicensed spectrum,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6310-6326, Sept. 2016.

[8] X. Wang, T. Q. S. Quek, M. Sheng and J. Li, “Throughput and fairness analysis of Wi-Fi and LTE-U in unlicensed band,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 1, pp. 63-78, Jan. 2017.

[9] C. Liu and H. Tsai, “On the limits of coexisting coverage and capacity in multi-RAT heterogeneous networks,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3086-3101, May 2017.

[10] N. Rastegardoost and B. Jabbari, “Minimizing Wi-Fi latency with unlicensed LTE opportunistic white-space utilization,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1914-1926, March 2019.

[11] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[12] R. Yin, G. Yu, A. Maaref, and G. Li, “A framework for co-channel interference and collision probability tradeoff in LTE licensed-assisted access networks,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6078–6090, Sept. 2016.

[13] S. Han, Y. C. Liang, Q. Chen and B. H. Soong, “Licensed-assisted access for LTE in unlicensed spectrum: A MAC protocol design,” *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2550–2561, Oct. 2016.

[14] Y. Song, K. W. Sung, and Y. Han, “Coexistence of Wi-Fi and cellular with listen-before-talk in unlicensed spectrum,” *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 161–164, Jan. 2016.

[15] Y. Ma and D. G. Kuester, “MAC-layer coexistence analysis of LTE and WLAN systems via listen-before-talk,” in *Proc. IEEE CCNC*, Las Vegas, NV, 2017, pp. 534-541.

[16] Y. Ma, D. G. Kuester, J. Coder and W. F. Young, “Slot-jamming effect and mitigation between LTE-LAA and WLAN systems with heterogeneous slot durations,” *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4407-4422, June 2019.

[17] Y. Ma and J. Coder, “Analysis of generalized CCA errors and mitigation in LTE-LAA spectrum sharing system,” *Proc. IEEE GlobeCom*, Hawaii, USA, Dec. 2019, pp. 1-7.

[18] A. Goldsmith, *Wireless Communications,* Cambridge University Press, 2005.

Penetration Loss at 60 GHz for Indoor-to-Indoor and Outdoor-to-Indoor Mobile Scenarios*

Sung Yun Jun[1], Derek Caudill[1], Jack Chuang[2], Peter B. Papazian[2], Anuraag Bodi[2], Camillo Gentile[2], Jelena Senic[2], Nada Golmie[2]

[1] RF Technology Division, National Institute of Standards and Technology, Boulder, CO, USA, sungyun.jun@nist.gov

[2] Wireless Networks Division, National Institute of Standards and Technology, Gaithersburg, MD, USA,

***Abstract*—This paper investigates the penetration loss of an office building in indoor-to-indoor (I2I) and outdoor-to-indoor (O2I) mobile scenarios. The measurements were collected using our 60 GHz double-directional switched-antenna channel sounder. During measurements, the transmitter, mounted on a tripod, was placed in an office and outside of the building, while the receiver, mounted on a mobile robot, moved along an interior hallway. The penetration loss for a variety of building materials was predicted versus incident angle by electromagnetic propagation theory using the International Telecommunication Union Radiocommunication Sector (ITU-R) Recommendation P.2040 model parameters and compared with the measurement results. The wooden door, plasterboard wall, and interior glass were observed to have penetration losses ranging from 25.5 dB to 40.5 dB, 11.8 dB to 31.6 dB, and 7.5 dB to 18.1 dB, respectively, while the exterior building materials exhibited even larger penetration losses, ranging from 31.1 dB to 66.5 dB.**

***Index Terms*—channel sounder, channel propagation model, 5G wireless communications, millimeter wave**

I. Introduction

Fifth generation (5G) wireless communications have created a growing demand for millimeter-wave (mmWave) channel sounding and modeling. The modeling and optimization of 5G cellular network technologies are highly dependent on the radio-wave propagation characteristics [1]. mmWave communications will operate below ideal capacity in non-line-of-sight conditions as a result of the properties of the obstructing materials that cause attenuation due to absorption and dispersion of the electromagnetic waves.

To date, several studies have been conducted to measure the signal attenuation of materials at mmWave frequencies and to model path loss in indoor and outdoor environments: Extensive measurements for the penetration loss of various materials at 28 GHz and 73 GHz in an indoor office environment were studied in [2]-[3]. An experimental investigation of a 60 GHz wireless local-area network system in an indoor cubicle environment was performed in [4]; the measurement campaigns were conducted using horn antennas at fixed points within an indoor environment; the study resulted in an empirical O2I building model for penetration loss. Analysis of reflection and penetration losses for common building materials at 28 GHz was presented in an urban outdoor environment [5]. The suburban residential neighborhood penetration loss at 28 GHz was also investigated in [6]. Wideband channel measurements in downtown Denver, CO at 9.6 GHz, 28.8 GHz and 57.6 GHz were collected in [7]; significant penetration loss was found to be caused by obstruction from an office building.

In general, previous literature dealing with penetration loss measurements has focused on stationary points using directional horn antennas with limited scan angles. Lacking to date is the effect on penetration loss of different points of incidence and different angles of incidence across the surface of building materials, to capture what a real communications system would experience during operation. To fill that void, this paper presents measured penetration loss at mmWave using an electronically switched double-directional channel sounder that was developed by the National Institute of Standards and Technology (NIST) [8]. The sounder has at the center frequency of 60.5 GHz: 2 GHz bandwidth, wide scan angle, and a mobile robot positioning platform. The measurement campaign was conducted in I2I and O2I environments. For the purpose of comparison, a theoretical analysis for the penetration loss of the building materials and its dependence on incident angle was also conducted.

The paper is organized as follows: The theoretical analysis and measurement setup are presented in Section II and Section III, respectively. Analysis of the penetration loss measurement results are described in Section IV, followed by conclusions in Section V.

II. Theoretical Analysis

Radio propagation at an interface with dielectric materials will depend on a number of parameters, most importantly the center frequency of the signal, the angle of incidence of the signal with the interface, and the material properties themselves. Generally speaking, transmission through materials degrade at higher frequencies, at shallower angles, and for denser materials. In order to provide a benchmark for comparison with our ensuing measurements, we first conducted a theoretical analysis of penetration loss by means of the Fresnel equations [9], which provide the reflection and transmission coefficients of electromagnetic waves incident on a flat surface. The material properties are represented through a complex relative permittivity η, expressed through real and imaginary components as

$$\eta = \eta' - j\eta''. \tag{1}$$

Jun, Sung Yun; Caudill, Derek; Chuang, Jack; Papazian, Peter; Bodi, Anuraag; Gentile, Camillo; Senic, Jelena; Golmie, Nada T. ”Penetration Loss at 60 GHz for Indoor-to-Indoor and Outdoor-to-Indoor Mobile Scenarios.” Paper presented at IEEE European Conference on Antennas and Propagation, Copenhagen, DK. March 15, 2020 - March 20, 2020.

TABLE I. MATERIAL PROPERTIES AT 60.5 GHZ

Material Class	Real part of relative permittivity (η')	Imaginary part of relative permittivity (η'')
Air	1	0
Metal	1	2.9719×10^{6}
Wood	1.99	0.1135
Plasterboard	2.94	0.0628
Glass	6.27	0.1703

The values of the frequency-dependent components were accordingly set to the center frequency of our channel sounder.

Table I shows the complex relative permittivity of several common construction materials at 60.5 GHz per the ITU-R Recommendation P.2040 [10]. We focused our theoretical analysis on wood, glass, and plasterboard only, the dominant materials in the environment where the measurements were collected. The thickness of the single-layered wood and glass materials was set as 45 mm and 9.5 mm, respectively, to match properties of the office building under investigation; the multi-layered wall, rather, was set to 12.7 mm for the two plasterboard sheets separated by an air pocket of 88.9 mm; in reality, as observed through the measurements, there were also metal studs in between the sheets.

Fig. 1 shows the theoretical attenuation of the three materials versus incident angle from 0° to 85°. The attenuation grows exponentially with incident angle. The wood and glass have attenuation in the range from 20.1 dB to 38.7 dB and from 5.8 dB to 19.9 dB, respectively. On the other hand, the multi-layer plasterboard attenuation varied from 5.5 dB to 32.3 dB over the same angle range. Interestingly, the ripple observed in the attenuation of the plasterboard can be traced to the multi-layered structure.

III. MEASUREMENTS

This section describes the channel sounder and the measurement campaign used to collect data for the I2I and O2I mobile scenarios.

A. Channel Sounder

Our 60 GHz double-directional switched array channel sounder is described in detail in [8]. The correlation-based system utilizes a pseudo-random noise (PN) sequence as the probing signal. For increased dynamic range to deal with greater range and greater penetration loss, a longer PN sequence of 32767 chips with chip rate of 2 Gbits/s was employed for the I2I scenario and the O2I scenario (at the expense of channel sweep time).

The transmitter (TX) is comprised of an intermediate frequency (IF) section, a radio frequency (RF) up-conversion section, and an eight-element antenna array with a switching multiplexer (MUX) (see Fig. 2(a,b)). The receiver (RX) contains an analog-to-digital converter section, an RF down-conversion section, and a sixteen-element antenna array with a MUX (see Fig. 2(c,d)). Both arrays contain scalar feed horns with 18.1 dBi gain and half-power beamwidth of 22.5°. Given the constellation of the elements, the synthesized antenna patterns of the TX and RX arrays provide 180° and 360° field-of-view (FoV), respectively, in the azimuthal plane whereas the FoV in the elevation plane is 45° for both.

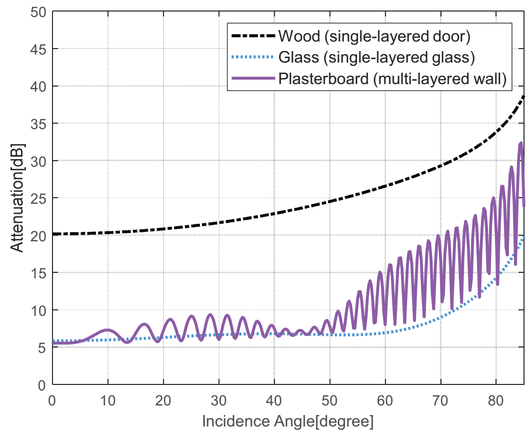
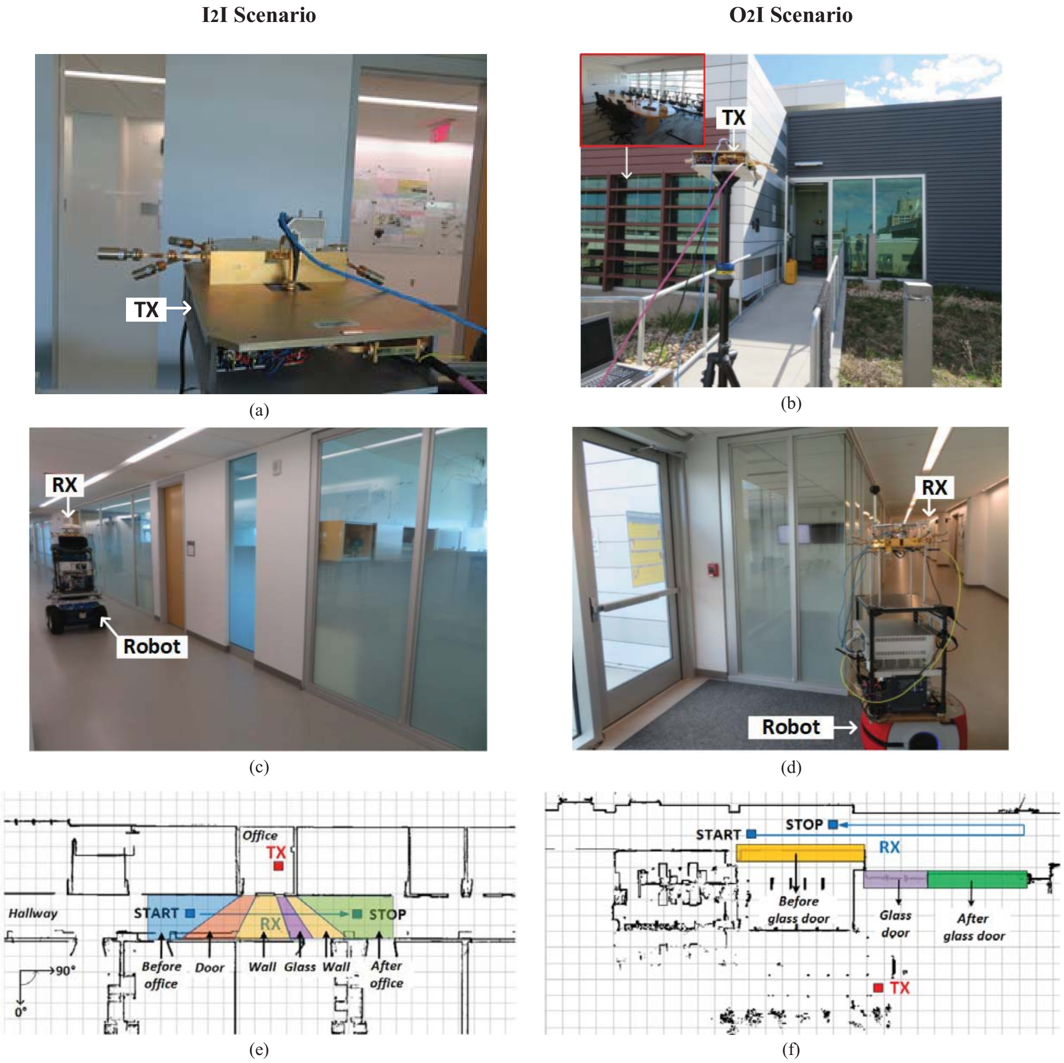


Fig. 1. Calculated transmission attenuation versus incidence angle for Wood, Plasterboard and Glass at 60.5 GHz.

Two 10 MHz Rubidium time standards and timing control circuits are used for untethered timing synchronization between the TX and RX for switching transmission and reception of the PN sequences between the 8×16=128 channels. An arbitrary waveform generator (AWG) at the TX generates the PN sequence, which is modulated using Binary Phase Shift Keying at an IF frequency of 3 GHz and then up-converted to an RF frequency of precisely 60.5 GHz. The TX antenna array transmits the RF signal with an equivalent isotropic radiated power (EIRP) of 36 dBm. The received signal at the RX antenna array is then down-converted to the IF of 3 GHz and is digitized at 40 Gsamples/s per channel. Matched filtering of the digitized signal with the PN sequence to generate the channel impulse response for each TX-RX antenna pair is performed off-line to decrease the channel sweep time. The resultant data for one TX-RX measurement point therefore consists of 128 channel impulse responses.

To remove the systematic distortion effects caused by the system hardware, a back-to-back calibration method was applied, as described in [8]. The calibration significantly reduces distortions and internal reflections of the system, extending the dynamic range to 90 dB for the longer PN sequence. The antenna patterns of the directional horns were characterized in an anechoic environment and de-embedded from the measurements as part of the calibration procedures. The 128 channel impulse responses recorded are post-processed through the Space Alternating Generalized Expectation maximization (SAGE) algorithm [11] to extract the channel multipath components (MPCs) and their properties, namely the delay, angle-of-departure (AoD) and angle-of-arrival (AoA) (in both azimuth and elevation), and



I2I Scenario **O2I Scenario**

(a) (b)

(c) (d)

(e) (f)

Fig. 2. Measurement setups for the I2I (left column) and O2I (right column) mobile scenarios. (a,b) TX positions (c,d) RX positions (e,f) Maps.

the path loss of each path identified. The aforementioned calibration procedures ensured that the features of the channel sounder were decoupled from the measurements such that properties of the extracted MPCs reflected the channel alone (and not the system).

B. *Measurement Setup*

The environment for the measurement campaign was a modern office building on the campus of the NIST in Boulder, CO, USA. Fig. 2(a,c) and Fig. 2(b,d) display photographs of the measurement setups for the I2I and O2I scenarios, respectively, and Fig. 2(e,f) show their maps. The maps were automatically generated by the laser-guided navigational system of the robot, furnishing precision localization to within a centimeter. For the I2I scenario, the fixed TX was mounted on a tripod inside an office at height 1.6 m and facing the adjacent hallway; the RX, mounted on a mobile robot, traversed the route shown in Fig. 2(e) from start to stop in the hallway, over which 57 data points were collected within the TX-RX distance range of 2.7 m to 6.1 m. For the O2I scenario, the TX was raised to 2.5 m and placed outside the building, still facing the hallway; the RX traversed a longer route in the same hallway, shown in Fig. 2(f), over which 93 data points were collected within the distance range of 9.4 m to 13.7 m. For both scenarios, the direct path between the TX and RX was obstructed by single- and multiple-layered materials.

Jun, Sung Yun; Caudill, Derek; Chuang, Jack; Papazian, Peter; Bodi, Anuraag; Gentile, Camillo; Senic, Jelena; Golmie, Nada T. "Penetration Loss at 60 GHz for Indoor-to-Indoor and Outdoor-to-Indoor Mobile Scenarios." Paper presented at IEEE European Conference on Antennas and Propagation, Copenhagen, DK. March 15, 2020 - March 20, 2020.

IV. Measurement Results

The penetration loss for each measurement point collected was computed as in [3]: First, the *penetrating path* was identified among all MPCs extracted per measurement point as the one that arrived first. Its delay (τ) was then mapped to the theoretical free-space path loss (PL_{FS}) through Friis transmission equation as

$$PL_{FS} = 20 \cdot log_{10}\left(\frac{4\pi c\tau}{\lambda}\right), \tag{2}$$

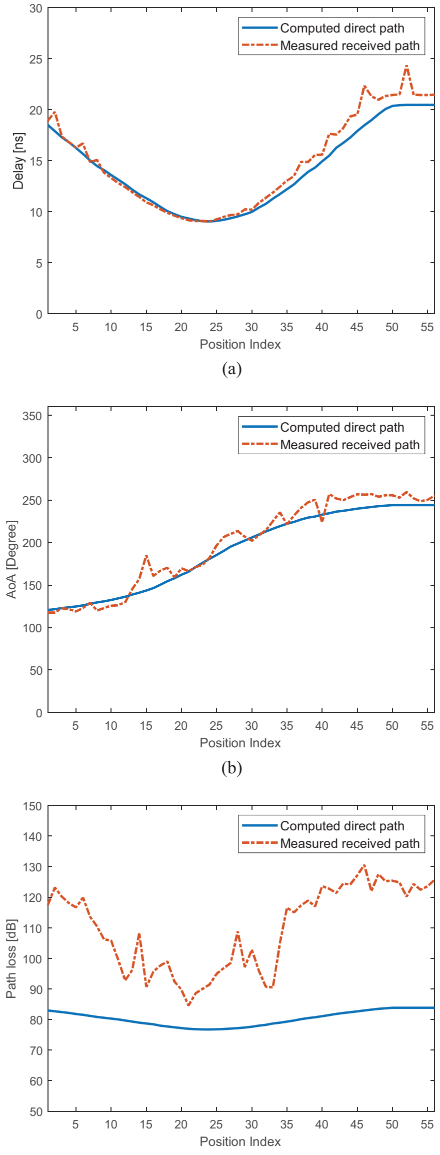
where c is the speed of light and $\lambda = 5$ mm is the wavelength corresponding to 60.5 GHz. Finally, the penetration loss was estimated by subtracting PL_{FS} from the measured path loss of the penetrating path. In the sequel, we present the estimated penetration loss for the I2I and O2I scenarios.

A. I2I Penetration Loss

For the purpose of verification, Fig. 3 displays the estimated delay, AoA, and path loss of the penetrating path (dashed red) versus the position index of the I2I scenario. Also displayed are the analogous theoretical values for the free-space direct path (solid blue), where its delay (τ_{DP}) and AoA were computed from the geometry of the known TX and RX positions and direct path loss (PL_{DP}) was computed by substituting τ_{DP} into (2). Aside from any system estimation error, the reason the estimated delay / AoA differ from the theoretical values is because the trajectory of the penetrating MPC deviated from the direct path due to reflection, refraction, and dispersion occurring at the material boundaries and/or in the environment; the difference in path loss, rather, is equivalent to the estimated penetration loss. Note that in general the deviation in delay / AoA increased with penetration loss.

In order to classify the estimated penetration loss according to the dominant materials in the environment, we partitioned the side of the hallway penetrated by the signal into *zones*; the partitioning was based on where the theoretical direct path intersected the side as the RX moved in the hallway. Fig. 2(e) displays the resultant partitioning with different colors: For the single penetration from the office, the side was constructed from a wooden door, plasterboard and glass sections; the measured penetration loss there ranged from 25.5 dB to 40.5 dB, from 11.8 dB to 31.6 dB, and from 7.5 dB to 18.1 dB, respectively. For the segments before and after the office, on the other hand, the side was constructed totally from plasterboard, through which multiple penetrations (including the adjacent vertical office walls) occurred; there the penetration loss increased in range from 34.4 dB to 40.5 dB and from 26.0 dB and 47.6 dB, respectively, as shown in Fig. 4(a). The interior glass featured the lowest penetration loss among the environment materials.

The penetration loss of the wooden door, plasterboard wall, and interior glass were slightly higher than the attenuation characteristics of the ITU-R P.2040 model (Fig. 1). It was determined that the door is made of fire-rated wood with a composite mineral core containing several different materials. The composite materials have more complex reflection and transmission effects, which cause the high attenuation versus the homogenous materials. Moreover, the multiple peaks of the penetration loss in the wall can be explained by the metal studs installed at regular spacing between the plasterboard sheets. The thin window film on the glass (for privacy) also provides a slightly higher penetration loss compared to clear glass (Fig. 1).



(a)

(b)

(c)

Fig. 3. Properties of the penetrating path identified from the I2I measurement versus the theoretical properties of the direct path. (a) delay (b) AoA and (c) path loss.

B. O2I Penetration Loss

Fig. 2(f) displays how the side of the hallway penetrated by the signal was partitioned based on the exterior construction materials for the O2I scenario, and Fig. 4(b) displays the corresponding results for the penetration loss estimated. Along the segment before the glass door with double-pane glass, there is the conference room between the TX and RX for which the penetrating path most likely went through an exterior double-pane window to the conference room and then out a single-pane interior window to the hallway. The exterior window consisted of a single, clear inner glass pane and a single, tinted outer glass pane separated by 19 mm (the outer pane tint had low emissivity coating applied to reject ultraviolet light). The interior windows were clear single pane glass. The segment after the glass door was an exterior wall. The segments before and after the glass door exhibited huge penetration losses, ranging between 52.3 dB to 66.5 dB and 40.3 dB to 59.2 dB, respectively. On the other hand, the glass door had a penetration loss in the range of 31.1 dB to 46.0 dB with an average value of 38.5 dB. It was observed that external building materials such as metal plating and low emissivity coatings caused high attenuation.

V. Conclusion

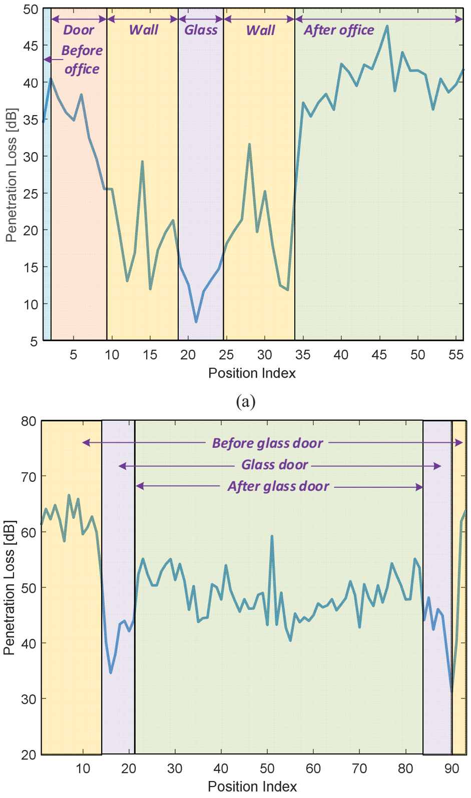
This paper presents measurements for penetration loss taken with our 60 GHz channel sounder over a variety of building materials in I2I and O2I mobile scenarios. Departing from previous efforts, the penetration loss was measured across continuous incident points and, by association, different incident angles along the surface of the building materials in the environment, to capture what an actual radio would experience in motion as opposed to a collection of separate, fixed-position measurements. As a means for comparison, the penetration loss for the building materials in the environment were modeled and analyzed theoretically through the Fresnel equations fit with the ITU-R P.2040 material parameters. The penetration loss through multiple layers was observed to be as high as 47.6 dB for I2I scenario, yet not as high as the 66.5 dB observed for the O2I scenario. In summary, we conclude that signals transmitted at 60 GHz both inside buildings and into buildings will suffer from severe penetration losses due to ambient building materials, heavily reducing coverage with respect to line-of-sight conditions. On the other hand, the high penetration loss of building materials can be a potential advantage to reduce interference from neighboring wireless network systems.

References

[1] K. A. Remley et al., "Measurement Challenges for 5G and Beyond: An Update from the National Institute of Standards and Technology," *IEEE Microwave Mag.*, vol. 18, no. 5, July– Aug. 2017, pp. 41–56.

[2] M. Lei, J. Zhang, T. Lei, and D. Du, "28-GHz indoor channel measurements and analysis of propagation characteristics," *Proc. IEEE Int. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, 2014, pp. 208–212.

[3] Ryan, J., MacCartney, G. R., and Rappaport, T., "Indoor Office Wideband Penetration Loss Measurements at 73 GHz", *IEEE International Conference on Communications Workshop (ICCW)*, pp.1-6, Paris, France, May 2017.



(a)

(b)

Fig. 4. Measured penetration loss for (a) I2I and (b) O2I scenarios.

[4] A. Maltsev, R. Maslennikov, A. Sevastyanov, A. Khoryaev, and A. Lomayev, "Experimental investigations of 60 GHz WLAN systems in office environment," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 8, pp. 1488–1499, Oct. 2009.

[5] H. Zhao et al., ''28 GHz millimeter wave cellular communication measurements for reflection and penetration loss in and around buildings in New York City,'' *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2013, pp. 5163–5167.

[6] J. Du, D. Chizhik, R. Feick, G. Castro, M. Rodr´ıguez and R. A. Valenzuela, "Suburban residential building penetration loss at 28 GHz for fixed wireless access," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 890–893, Dec. 2018.

[7] E. J. Violette et al., "Millimeter-wave propagation at street level in an urban environment," *IEEE Trans. Geosci. Remote Sens.*, vol. 26, no. 3, pp. 368–380, 1988.

[8] R. Sun, P. B. Papazian, J. Senic, Y. Lo, J.-K. Choi, K. A. Remley, and C. Gentile, "Design and calibration of a double-directional 60 GHz channel sounder for multipath component tracking," *Proc. IEEE European Conf. Antennas and Propagation*, pp. 1–5, Mar. 2017.

[9] J. Medbo and S. Dwivedi, "Frequency Selectivity of Window Attenuation up to 100 GHz," *2019 13th European Conference on Antennas and Propagation (EuCAP)*, Krakow, Poland, 2019, pp. 1-4.

[10] Recommendation ITU-R P.2040, "Effects of building materials and structures on radiowave propagation above about 100 MHz," https://www.itu.int/rec/R-REC-P.2040-1-201507-I/en.I.S.

[11] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Process.,* vol. 42, no. 10, pp. 2664-2677, Oct. 1994.

Discovering Mathematical Objects of Interest—A Study of Mathematical Notations

André Greiner-Petter
University of Wuppertal
Germany
andre.greiner-petter@zbmath.org

Moritz Schubotz
FIZ-Karlsruhe and
University of Wuppertal
Germany
moritz.schubotz@fiz-karlsruhe.de

Fabian Müller
FIZ-Karlsruhe
Germany
fabian.mueller@fiz-karlsruhe.de

Corinna Breitinger
University of Wuppertal and
University of Konstanz
Germany
corinna.breitinger@uni-konstanz.de

Howard S. Cohl
Applied and Computational
Mathematics Division, National
Institute of Standards and Technology,
Mission Viejo, California, U.S.A.
howard.cohl@nist.gov

Akiko Aizawa
National Institute of Informatics
Japan
aizawa@nii.ac.jp

Bela Gipp
University of Wuppertal and
University of Konstanz
Germany
gipp@uni-wuppertal.de

ABSTRACT

Mathematical notation, i.e., the writing system used to communicate concepts in mathematics, encodes valuable information for a variety of information search and retrieval systems. Yet, mathematical notations remain mostly unutilized by today's systems. In this paper, we present the first in-depth study on the distributions of mathematical notation in two large scientific corpora: the open access arXiv (2.5B mathematical objects) and the mathematical reviewing service for pure and applied mathematics zbMATH (61M mathematical objects). Our study lays a foundation for future research projects on mathematical information retrieval for large scientific corpora. Further, we demonstrate the relevance of our results to a variety of use-cases. For example, to assist semantic extraction systems, to improve scientific search engines, and to facilitate specialized math recommendation systems.

The contributions of our presented research are as follows: (1) we present the first distributional analysis of mathematical formulae on arXiv and zbMATH; (2) we retrieve relevant mathematical objects for given textual search queries (e.g., linking $P_n^{(\alpha,\beta)}(x)$ with 'Jacobi polynomial'); (3) we extend zbMATH's search engine by providing relevant mathematical formulae; and (4) we exemplify the applicability of the results by presenting auto-completion for math inputs as the first contribution to math recommendation systems. To expedite future research projects, we have made available our source code and data.

CCS CONCEPTS

• **Information systems → Mathematics retrieval;** • **Information systems → Novelty in information retrieval;** *Information extraction*; • **Information systems~Recommender systems;** • **Information systems~Near-duplicate and plagiarism detection**;

KEYWORDS

Mathematical Objects of Interest, Mathematical Information Retrieval, Distributions of Mathematical Objects, Term Frequency-Inverse Document Frequency, Mathematical Search Engine

ACM Reference Format:
André Greiner-Petter, Moritz Schubotz, Fabian Müller, Corinna Breitinger, Howard S. Cohl, Akiko Aizawa, and Bela Gipp. 2020. Discovering Mathematical Objects of Interest—A Study of Mathematical Notations. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3366423.3380218

1 INTRODUCTION

Taking into account mathematical notation in the literature leads to a better understanding of scientific literature on the Web and allows one to make use of semantic information in specialized Information Retrieval (IR) systems. Nowadays applications in Math Information Retrieval (MathIR) [15], such as search engines [7, 17, 18, 22, 24, 27, 31], semantic extraction systems [23, 36, 37], recent efforts in math embeddings [10, 13, 25, 44], and semantic tagging of math formulae [6, 43] either consider an entire equation as one entity or only focus on single symbols. Since math expressions often contain meaningful and important subexpressions, these applications could benefit from an approach that lies between the extremes of examining only individual symbols or considering an entire equation as one entity. Consider for example, the explicit definition for Jacobi polynomials [8, (18.5.7)]

$$P_n^{(\alpha,\beta)}(x)=\frac{\Gamma(\alpha+n+1)}{n!\Gamma(\alpha+\beta+n+1)}\sum_{m=0}^{n}\binom{n}{m}\frac{\Gamma(\alpha+\beta+n+m+1)}{\Gamma(\alpha+m+1)}\left(\frac{x-1}{2}\right)^m. \quad (1)$$

Cohl, Howard; Greiner Petter, Andre; Schubotz, Moritz; Breitinger, Corinna; Muller, Fabian; Aizawa, Akiko; Gipp, Bela. "Discovering Mathematical Objects of Interest - A Study of Mathematical Notations." Paper presented at The Web Conference 2020, Taipei, TW. April 20, 2020 - April 24, 2020.

The *interesting* components in this equation are $P_n^{(\alpha,\beta)}(x)$ on the left-hand side, and the appearance of the gamma function $\Gamma(s)$ on the right-hand side, implying a direct relationship between Jacobi polynomials and the gamma function. Considering the entire expression as a single object misses this important relationship. On the other hand, focusing on single symbols can result in the misleading interpretation of Γ as a variable and $\Gamma(\alpha + n + 1)$ as a multiplication between Γ and $(\alpha + n + 1)$. A system capable of identifying the important components, such as $P_n^{(\alpha,\beta)}(x)$ or $\Gamma(\alpha + n+1)$, is therefore desirable. Hereafter, we define these components as *Mathematical Objects of Interest* (MOIs) [13].

The *importance* of math objects is a somewhat imprecise description and thus difficult to measure. Currently, not much effort has been made in identifying meaningful subexpressions. Kristianto et al. [23] introduced dependency graphs between formulae. With this approach, they were able to build dependency graphs of mathematical expressions, but only if the expressions appeared as single expressions in the context. For example, if $\Gamma(\alpha + n + 1)$ appears as a stand-alone expression in the context, the algorithm will declare a dependency with Equation (1). However, it is more likely that different forms, such as $\Gamma(s)$, appear in the context. Since this expression does not match any subexpression in Equation (1), the approach cannot establish a connection with $\Gamma(s)$. Kohlhase et al. studied in [19–21] another approach to identify essential components in formulae. They performed eye-tracking studies to identify important areas in rendered mathematical formulae. While this is an interesting approach that allows one to learn more about the insights of human behaviors of reading and understanding math, it is inaccessible for extensive studies.

This paper presents the first extensive frequency distribution study of mathematical equations in two large scientific corpora, the e-Print archive arXiv.org (hereafter referred to as arXiv[1]) and the international reviewing service for pure and applied mathematics zbMATH[2]. We will show that math expressions, similar to words in natural language corpora, also obey Zipf's law [33], and therefore follows a *Zipfian* distribution. Related research projects observed a relation to Zipf's law for single math symbols [6, 36]. In the context of quantitative linguistics, Zipf's law states that given a text corpus, the frequency of any word is inversely proportional to its rank in the frequency table. Motivated by the similarity to linguistic properties, we will present a novel approach for ranking formulae by their relevance via a customized version of the ranking function BM25 [34]. We will present results that can be easily embedded in other systems in order to distinguish between common and uncommon notations within formulae. Our results lay a foundation for future research projects in MathIR.

Fundamental knowledge on frequency distributions of math formulae is beneficial for numerous applications in MathIR, ranging from educational purposes [40] to math recommendation systems, search engines [7, 31], and even automatic plagiarism detection systems [28, 29, 39]. For example, students can search for the conventions to write certain quantities in formulae; document preparation systems can integrate an auto-completion or auto-correction service for math inputs; search or recommendation engines can adjust their ranking scores with respect to standard notations; and plagiarism detection systems can estimate whether two identical formulae indicate potential plagiarism or are just using the conventional notations in a particular subject area. To exemplify the applicability of our findings, we present a textual search approach to retrieve mathematical formulae. Further, we will extend zbMATH's faceted search by providing facets of mathematical formulae according to a given textual search query. Lastly, we present a simple auto-completion system for math inputs as a contribution towards advancing mathematical recommendation systems. Further, we show that the results provide useful insights for plagiarism detection algorithms. We provide access to the source code, the results, and extended versions of all of the figures appearing in this paper at https://github.com/ag-gipp/FormulaCloudData.

Related Work: Today, mathematical search engines index formulae in a database. Much effort has been undertaken to make this process as efficient as possible in terms of precision and runtime performance [7, 17, 26, 27, 45]. The generated databases naturally contain the information required to examine the distributions of the indexed mathematical formulae. Yet, no in-depth studies of these distributions have been undertaken. Instead, math search engines focus on other aspects, such as devising novel similarity measures and improving runtime efficiency. This is because the goal of math search engines is to retrieve relevant (i.e., similar) formulae which correspond to a given search query that partially [24, 26, 31] or exclusively [7, 17, 18] contains formulae. However, for a fundamental study of distributions of mathematical expressions, no similarity measures nor efficient lookup or indexing is required. Thus, we use the general-purpose query language XQuery and employ the BaseX[3] implementation. BaseX is a free open-source XML database engine, which is fully compatible with the latest XQuery standard [14, 41]. Since our implementations rely on XQuery, we are able to switch to any other database which allows for processing via XQuery.

2 DATA PREPARATION

LaTeX is the de facto standard for the preparation of academic manuscripts in the fields of mathematics and physics [11]. Since LaTeX allows for advanced customizations and even computations, it is challenging to process. For this reason, LaTeX expressions are unsuitable for an extensive distribution analysis of mathematical notations. For mathematical expressions on the web, the XML formatted MathML[4] is the current standard, as specified by the World Wide Web Consortium (W3C). The tree structure and the fixed standard, i.e., MathML tags, cannot be changed, thus making this data format reliable. Several available tools are able to convert from LaTeX to MathML [35] and various databases are able to index XML data. Thus, for this study, we have chosen to focus on MathML. In the following, we investigate the databases arXMLiv (08/2018) [12] and zbMATH[5] [38].

The arXMLiv dataset (≈1.2 million documents) contains HTML5 versions of the documents from the e-Print archive arXiv.org. The HTML5 documents were generated from the TeX sources

[1] https://arxiv.org/ [Accessed: Sep. 1, 2019]
[2] https://zbmath.org [Accessed: Sep. 1, 2019]
[3] http://basex.org/ [Accessed: Sep. 2019]; We used BaseX 9.2 for our experiments.
[4] https://www.w3.org/TR/MathML3/ [Accessed: Sep. 1, 2019]
[5] https://zbmath.org/ [Accessed: Sep. 1, 2019]

via LaTeXML [30]. LaTeXML converted all mathematical expressions into MathML with parallel markup, i.e., presentation and content MathML. In this study we only consider the subsets *no-problem* and *warning*, which generated no errors during the conversion process. Nonetheless, the MathML data generated still contains some errors or falsely annotated math. For example, we discovered several instances of affiliation and footnotes, SVG[6] and other unknown tags, encoded in MathML. Regarding the footnotes, we presumed that authors falsely used mathematical environments for generating footnote or affiliation marks. We used the TeX string, provided as an attribute in the MathML data, to filter out expressions that match the string '`{}^{*}`', where '`*`' indicates any possible expression. In addition, we filtered out SVG and other unknown tags. We assume that these expressions were generated by mistake due to limitations of LaTeXML. The final arXiv dataset consisted of 841,008 documents which contained at least one mathematical formula. The dataset contained a total of 294,151,288 mathematical expressions.

In addition to arXiv, we investigated zbMATH, an international reviewing service for pure and applied mathematics which contains abstracts and reviews of articles, hereafter uniformly called abstracts, mainly from the domains of pure and applied mathematics. The abstracts in zbMATH are formatted in TeX [38]. To be able to compare arXiv and zbMATH, we manually generated MathML via LaTeXML for each mathematical formula in zbMATH and performed the same filters as used for the arXiv documents. The zbMATH dataset contained 2,813,451 abstracts, of which 1,349,297 contained at least one formula. In total, the dataset contained 11,747,860 formulae. Even though the total number of formulae is smaller compared to arXiv, we hypothesize that math formulae in abstracts are particularly meaningful.

2.1 Data Wrangling

Since we focused on the frequency distributions of visual expressions, we only considered presentational MathML (pMML). Rather than normalizing the pMML data, e.g., via MathMLCan [9], which would also change the tree structure and visual core elements in pMML, we only eliminated the attributes. These attributes are used for minor visual changes, e.g., stretched parentheses or inline limits of sums and integrals. Thus, for this first study, we preserved the core structure of the pMML data, which might provide insightful statistics for the MathML community to further cultivate the standard. After extracting all MathML expressions, filtering out falsely annotated math and SVG tags, and eliminating unnecessary attributes and annotations, the datasets required 83GB of disk space for arXiv and 6GB for zbMATH, respectively.

Listing 1: MathML representation of $P_n^{(\alpha,\beta)}(x)$.

```
 1 <math><mrow>
 2  <msubsup>
 3   <mi>P</mi>
 4   <mi>n</mi>
 5   <mrow>
 6    <mo>(</mo>
 7    <mi>α</mi>
 8    <mo>,</mo>
 9    <mi>β</mi>
10    <mo>)</mo>
11    <mo></mo>
12   </mrow>
13  </msubsup>
14  <mo></mo>
15  <mrow>
16   <mo>(</mo>
17   <mi>x</mi>
18   <mo>)</mo>
19  </mrow>
20 </mrow></math>
```

In the following, we indexed the data via BaseX. The indexed datasets required a disk space of 143.9GB in total (140GB for arXiv and 3.9GB for zbMATH). Due to the limitations[7] of databases in BaseX, it was necessary to split our datasets into smaller subsets. We split the datasets according to the 20 major article categories of arXiv[8] and classifications of zbMATH. To increase performance, we use BaseX in a server-client environment. We experienced performance issues in BaseX when multiple clients repeatedly requested data from the same server in short intervals. We determined that the best workaround for this issue was to launch BaseX servers for each database, i.e., each category/classification.

Mathematical expressions often consist of multiple meaningful subexpressions, which we defined as MOIs. However, without further investigation of the context, it is impossible to determine meaningful subexpressions. As a consequence, every equation is a potential MOI on its own and potentially consists of multiple other MOIs. For an extensive frequency distributional analysis, we aim to discover all possible mathematical objects. Hence, we split every formula into its components. Since MathML is an XML data format (essentially a tree-structured format), we define subexpressions of equations as subtrees of its MathML format.

Listing 1 illustrates a Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ in pMML. The `<mo>` element on line 14 contains the *invisible times* UTF-8 character. By definition, the `<math>` element is the root element of MathML expressions. Since we cut off all other elements besides pMML nodes, each `<math>` element has one and only one child element[9]. Thus, we define the child element of the `<math>` element as the root of the expression. Starting from this root element, we explore all subexpressions. For this study, we presume that every meaningful mathematical object (i.e., MOI) must contain at least one identifier.

Hence, we only study subtrees which contain at least one `<mi>` node. Identifiers, in the sense of MathML, are '*symbolic names or arbitrary text*' [10], e.g., single Latin or Greek letters. Identifiers do not contain special characters (other than Greek letters) or numbers. As a consequence, arithmetic expressions, such as $(1+2)^2$, or sequences of special characters and numbers, such as $\{1, 2, \ldots\} \cap \{-1\}$, will not appear in our distributional analysis. However, if a sequence or arithmetic expression consists of an identifier somewhere in the pMML tree (such as in $\{1, 2, \ldots\} \cap A$), the entire expression will be recognized. The Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$, therefore consists of the following subexpressions: $P_n^{(\alpha,\beta)}$, (α, β), (x), and the single identifiers P, n, α, β, and x. The entire expression is also a mathematical object. Hence, we take entire expressions with an identifier into account for our analysis. In the following, the set of subexpressions will be understood to include the expression itself.

For our experiments, we also generated a string representation of the MathML data. The string is generated recursively by applying one of two rules for each node: (i) if the current node is a leaf, the node-tag and the content

[6] Scalable Vector Graphics

[7] A detailed overview of the limitations of BaseX databases can be found at http://docs.basex.org/wiki/Statistics [Accessed: Sep. 1, 2019].

[8] The arXiv categories *astro-ph* (astro physics), *cond-mat* (condensed matter), and *math* (mathematics) were still too large for a single database. Thus, we split those categories into two equally sized parts.

[9] Sequences are always nested in an `<mrow>` element.

[10] https://www.w3.org/TR/MathML3/chapter3.html [Accessed: Sep. 1, 2019]

will be merged by a colon, e.g., `<mi>x</mi>` will be converted to `mi:x`; (ii) otherwise the node-tag wraps parentheses around its content and separates the children by a comma, e.g., `<mrow><mo>(</mo><mi>x</mi><mo>)</mo></mrow>` will be converted to `mrow(mo:(,mi:x,mo:))`. Furthermore, the special UTF-8 characters for invisible times (U+2062) and function application (U+2061) are replaced by `ivt` and `fa`, respectively. For example, the gamma function with argument $x+1$, $\Gamma(x+1)$ would be represented by

$$\texttt{mrow(mi:}\Gamma\texttt{,mo:ivt,mrow(mo:(,mrow(mi:x,mo:+,mn:1),mo:)))}. \quad (2)$$

Between Γ and $(x+1)$, there would most likely be the special character for *invisible times* rather than for *function application*, because LaTeXML is not able to parse Γ as a function. Note that this string conversion is a bijective mapping. The string representation reduces the verbose XML format to a more concise presentation. Thus, an equivalence check between two expressions is more efficient.

2.2 Complexity of Math

Mathematical expressions can become complex and lengthy. The tree structure of MathML allows us to introduce a measure that reflects the complexity of mathematical expressions. More complex expressions usually consist of more extensively nested subtrees in the MathML data. Thus, we define the complexity of a mathematical expression by the maximum depth of the MathML tree. In XML the content of a node and its attributes are commonly interpreted as children of the node. Thus, we define the depth of a single node as 1 rather than 0, i.e., single identifiers, such as `<mi>P</mi>`, have a complexity of 1. The Jacobi polynomial from Listing 1 has a complexity of 4.

We perform the extraction of subexpressions from MathML in BaseX. The algorithm for the extraction process is written in XQuery. The algorithm traverses recursively downwards from the root to the leaves. In each iteration, it checks whether there is an identifier, i.e., `<mi>` element, among the descendants of the current node. If there is no such element, the subtree will be ignored. It seems counterintuitive to start from the root and check if an identifier is among the descendants rather than starting at each identifier and traversing upwards to the root. If an XQuery requests a node in BaseX, BaseX loads the entire subtree of the requested node into the cache (up to a specified size). If the algorithm traverses upwards through the MathML tree, the XQuery will trigger database requests in every iteration. Hence, the downwards implementation performs better, since there is only one database request for every expression rather than for every subexpression.

Since we only minimize the pMML data rather than normalizing it, two identically rendered expressions may have different complexities. For instance, `<mrow><mi>x</mi></mrow>` consists of two distinct subexpressions, but both of them are displayed the same. Another problem often appears for arrays or similar visually complicated structures. The extracted expressions are not necessarily logical subexpressions. We will consider applying more advanced embedding techniques such as special tokenizers [26], symbol layout trees [7, 45], and a MathML normalization via MathMLCan [9] in future research to overcome these issues.

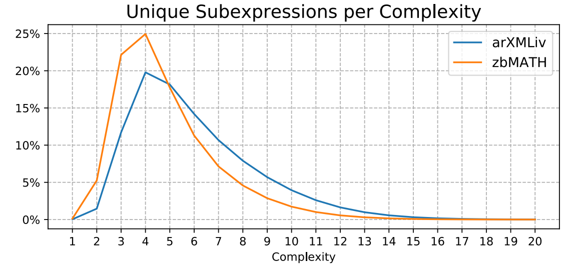


Figure 1: Unique subexpressions for each complexity in arXiv and zbMATH.

3 FREQUENCY DISTRIBUTIONS OF MATHEMATICAL FORMULAE

By splitting each formula into subexpressions, we generated longer documents and a bias towards low complexities. Note that, hereafter, we only refer to the mathematical content of documents. Thus, the length of a document refers to the number of math formulae—here the number of subexpressions—in the document. After splitting expressions into subexpressions, arXiv consists of 2.5B and zbMATH of 61M expressions, which raised the average document length to 2,982.87 for arXiv and 45.47 for zbMATH, respectively.

For calculating frequency distributions, we merged two subexpressions if their string representations were identical. Remember, the string representation is unique for each MathML tree. After merging, arXiv consisted of 350,206,974 unique mathematical subexpressions with a maximum complexity of 218 and an average complexity of 5.01. For high complexities over 70, the formulae show some erroneous structures that might be generated from LaTeXML by mistake. For example, the expression with the highest complexity is a long sequence of a polynomial starting with '$P_4(t_1, t_3, t_7, t_{11}) =$' followed by 690 summands. The complexity is caused by a high number of unnecessarily deeply nested `<mrow>` nodes. The highest complexity with a minimum document frequency of two is 39, which is a continued fraction. Since continued fractions are nested fractions, they naturally have a large complexity. One of the most complex expressions (complexity 20) with a minimum document frequency of three was the formula

$$\left(\sum_{j_1=1}^{n}\left(\sum_{j_2=1}^{n}\left(\cdots\left(\sum_{j_m=1}^{n}\left|T\left(e_{j_1},\ldots,e_{j_m}\right)\right|^{q_m}\right)^{\frac{q_{m-1}}{q_m}}\cdots\right)^{\frac{q_2}{q_3}}\right)^{\frac{q_1}{q_2}}\right)^{\frac{1}{q_1}} \leq C^{\mathbb{K}}_{m,p,\mathbf{q}}\|T\|. \quad (3)$$

In contrast, zbMATH only consisted of 8,450,496 unique expressions with a maximum complexity of 26 and an average complexity of 3.89. One of the most complex expressions in zbMATH with a minimum document frequency of three was

$$M_p(r,f) = \left(\frac{1}{2\pi}\int_0^{2\pi}\left|f\left(re^{i\theta}\right)\right|^p d\theta\right)^{1/p}. \quad (4)$$

As we expected, reviews and abstracts in zbMATH were generally shorter and consisted of less complex mathematical formulae. The

Category	arXiv	zbMATH
Documents	841,008	1,349,297
Formulae	294,151,288	11,747,860
Subexpressions	2,508,620,512	61,355,307
Unique Subexpressions	350,206,974	8,450,496
Average Document Length	2,982.87	45.47
Average Complexity	5.01	3.89
Maximum Complexity	218	26

Table 1: Dataset overview. Average Document Length is defined as the average number of subexpressions per document.

dataset also appeared to contain fewer erroneous expressions, since expressions of complexity 25 are still readable and meaningful.

Figure 1 shows the ratio of unique subexpressions for each complexity in both datasets. The figure illustrates that both datasets share a peak at complexity four. Compared to zbMATH, the arXiv expressions are slightly more evenly distributed over the different levels of complexities. Interestingly, complexities one and two are not dominant in either of the two datasets. Single identifiers only make up 0.03% in arXiv and 0.12% in zbMATH, which is comparable to expressions of complexity 19 and 14, respectively. This finding illustrates the problem of capturing semantic meanings for single identifiers rather than for more complex expressions [37]. It also substantiates that entire expressions, if too complex, are not suitable either for capturing the semantic meanings [23]. Instead, a middle ground is desirable, since the most unique expressions in both datasets have a complexity between 3 and 5. Table 1 summarizes the statistics of the examined datasets.

3.1 Zipf's Law

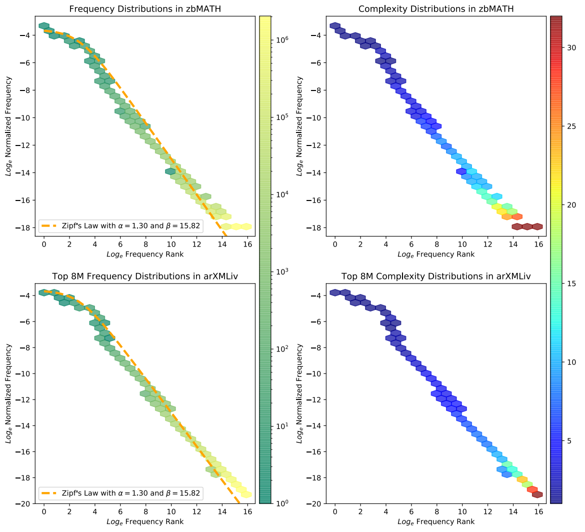
In linguistics, it is well known that word distributions follow Zipf's Law [33], i.e., the r-th most frequent word has a frequency that scales to

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{5}$$

with $\alpha \approx 1$. A better approximation can be applied by a shifted distribution

$$f(r) \propto \frac{1}{(r+\beta)^{\alpha}}, \tag{6}$$

where $\alpha \approx 1$ and $\beta \approx 2.7$. In a study on Zipf's law, Piantadosi [33] illustrated that not only words in natural language corpora follow this law surprisingly accurately, but also many other human-created sets. For instance, in programming languages, in biological systems, and even in music. Since mathematical communication has derived as the result of centuries of research, it would not be surprising if mathematical notations would also follow Zipf's law. The primary conclusion of the law illustrates that there are some very common tokens against a large number of symbols which are not used frequently. Based on this assumption, we can postulate that a score based on frequencies might be able to measure the peculiarity of a token. The infamous TF-IDF ranking functions and their derivatives [1, 34] have performed well in linguistics for many years and are still widely used in retrieval systems [3]. However, since we split every expression into its subexpressions, we generated an anomalous bias towards shorter, i.e., less complex, formulae. Hence, distributions of subexpressions may not obey Zipf's law.



(a) Frequency Distributions **(b) Complexity Distributions**

Figure 2: Each figure illustrates the relationship between the frequency ranks (x-axis) and the normalized frequency (y-axis) in zbMATH (top) and arXiv (bottom). For arXiv, only the first 8 million entries are plotted to be comparable with zbMATH (≈ 8.5 million entries). Subfigure (a) shades the hexagonal bins from green to yellow using a logarithmic scale according to the number of math expressions that fall into a bin. The dashed orange line represents Zipf's distribution (6). The values for α and β are provided in the plots. Subfigure (b) shades the bins from blue to red according to the maximum complexity in each bin.

Figure 2 visualizes a comparison between Zipf's law and the frequency distributions of mathematical subexpressions in arXiv and zbMATH. The dashed orange line visualizes the power law (6). The plots demonstrate that the distributions in both datasets obey this power law. Interestingly, there is not much difference in the distributions between both datasets. Both distributions seem to follow the same power law, with $\alpha = 1.3$ and $\beta = 15.82$. Moreover, we can observe that the developed complexity measure seems to be appropriate, since the complexity distributions for formulae are similar to the distributions for the length of words [33]. In other words, more complex formulae, as well as long words in natural languages, are generally more specialized and thus appear less frequent throughout the corpus. Note that colors of the bins for complexities fluctuate for rare expressions because the color represents the maximum rather than the average complexity in each bin.

3.2 Analyzing and Comparing Frequencies

Figure 3 shows in detail the most frequently used mathematical expressions in arXiv for the complexities 1 to 5. The orange dashed line visible in all graphs represents the normal Zipf's law distribution from Equation (5). We explore the total frequency values without

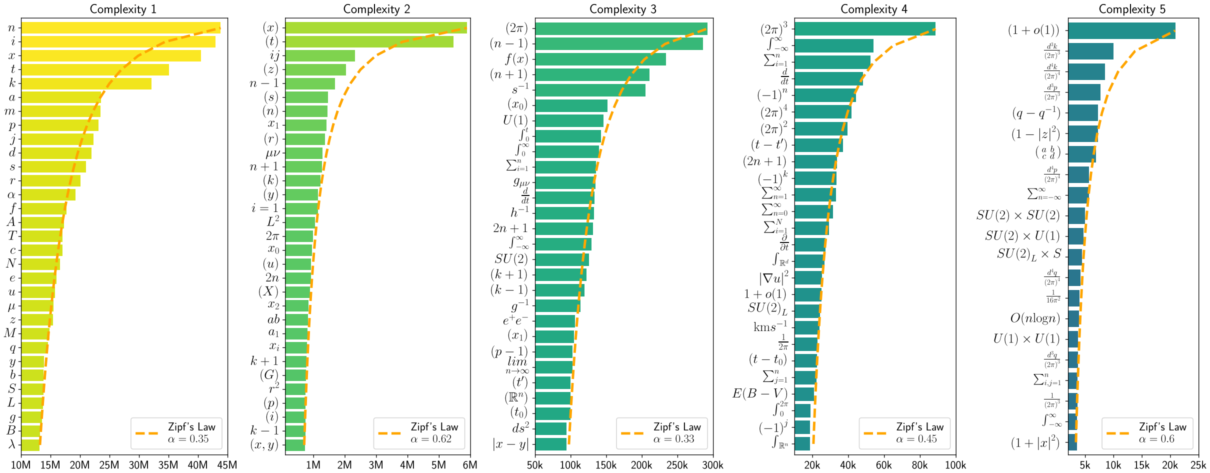


Figure 3: Overview of the most frequent mathematical expressions in arXiv for complexities 1-5. The color gradient from yellow to blue represents the frequency in the dataset. Zipf's law (5) is represented by a dashed orange line.

any normalization. Thus, Equation (5) was multiplied by the highest frequency for each complexity level to fit the distribution. The plots in Figure 3 demonstrate that even though the parameter α varies between 0.35 and 0.62, the distributions in each complexity class also obey Zipf's law.

The plots for each complexity class contain some interesting fluctuations. We can spot a set of five single identifiers that are most frequently used throughout arXiv: n, i, x, t, and k. Even though the distributions follow Zipf's law accurately, we can explore that these five identifiers are proportionally more frequently used than other identifiers and clearly separate themselves above the rest (notice the large gap from k to a). All of the five identifiers are known to be used in a large variety of scenarios. Surprisingly, one might expect that common pairs of identifiers would share comparable frequencies in the plots. However, typical pairs, such as x and y, or α and β, possess a large discrepancy.

The plot of complexity two also reveals that two expressions are proportionally more often used than others: (x) and (t). These two expressions appear more than three times as often in the corpus than any other expression of the same complexity. On the other hand, the quantitative difference between (x) and (t) is negligible. We may assume that arXiv's primary domain, physics, causes the quantitative disparity between (x), (t), and the other tokens. The primary domain of the dataset becomes more clearly visible for higher complexities, such as $SU(2)$ (C3[11]) or kms^{-1} (C4).

Another surprising property of arXiv is that symmetry groups, such as $SU(2)$, appear to play an essential role in the majority of articles on arXiv, see $SU(2)$ (C3), $SU(2)_L$ (C4), and $SU(2) \times SU(2)$ (C5), among others. The plots of higher complexities[12], which we do not show here, made this even more noticeable. Given a complexity of six, for example, the most frequently used expression was $SU(2)_L \times SU(2)_R$, and for a complexity of seven it was $SU(3) \times SU(2) \times U(1)$. Given a complexity of eight, ten out of the top-12 expressions were from symmetry group calculations.

It is also worthwhile to compare expressions among different levels of complexities. For instance, (x) and (t) appeared almost six million times in the corpus, but $f(x)$ (at position three in C3) was the only expression which contained one of these most common expressions. Note that subexpressions of variations, such as (x_0), (t_0), or $(t-t')$, do not match the expression of complexity two. This may imply that (x), and especially (t), appear in many different scenarios. Further, we can examine that even though (x) is a part of $f(x)$ in only approximately 3% of all cases, it is still the most likely combination. These results are especially useful for recommendation systems that make use of math as input. Moreover, plagiarism detection systems may also benefit from such a knowledge base. For instance, it might be evident that $f(x)$ is a very common expression, but for automatic systems that work on a large scale, it is not clear whether duplicate occurrences of $f(x)$ or $\Xi(x)$ should be scored differently, e.g., in the case of plagiarism detection.

Figure 3 shows only the most frequently occurring expressions in arXiv. Since we already explored a bias towards physics formulae in arXiv, it is worth comparing the expressions present within both datasets. Figure 4 compares the 25-top expressions for the complexities one to four. In zbMATH, we discovered that computer science and graph theory appeared as popular topics, see for example $G = (V, E)$ (in C3 at position 20) and the Bachmann-Landau notations in $O(\log n)$, $O(n^2)$, and $O(n^3)$ (C4 positions 2, 3, and 19).

From Figure 4, we can also deduce useful information for MathIR tasks which focus on semantic information. Current semantic extraction tools [37] or LaTeX parsers [35] still have difficulties distinguishing *multiplications* from *function calls*. For example as mentioned before, LaTeXML [30] adds an *invisible times* character

[11] We refer to a given complexity n with Cn, i.e., C3 refers to complexity 3.

[12] More plots showing higher complexities are available at https://github.com/ag-gipp/FormulaCloudData

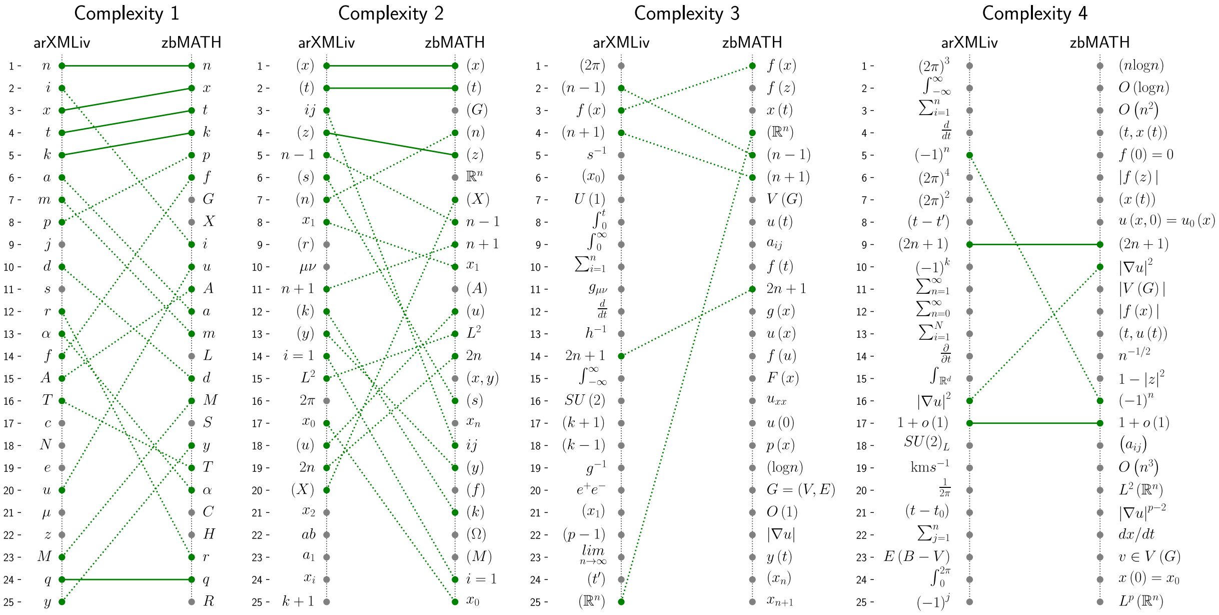

Complexity 1
arXMLiv
zbMATH
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 - 11 - 12 - 13 - 14 - 15 - 16 - 17 - 18 - 19 - 20 - 21 - 22 - 23 - 24 - 25 -
n i x t k a m p j d s r α f A T c N e u μ z M q y
n x t k p f G X i u A a m L d M S y T α C H r q R
Complexity 2
arXMLiv
zbMATH
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 - 11 - 12 - 13 - 14 - 15 - 16 - 17 - 18 - 19 - 20 - 21 - 22 - 23 - 24 - 25 -
(x) (t) ij (z) $n-1$ (s) (n) x_1 (r) $\mu\nu$ $n+1$ (k) (y) $i=1$ L^2 2π x_0 (u) $2n$ (X) x_2 ab a_1 x_i $k+1$
(x) (t) (G) (n) (z) $\mathbb{R}^n$ (X) $n-1$ $n+1$ x_1 (A) (u) L^2 $2n$ (x,y) (s) x_n ij (y) (f) (k) (Ω) (M) $i=1$ x_0
Complexity 3
arXMLiv
zbMATH
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 - 11 - 12 - 13 - 14 - 15 - 16 - 17 - 18 - 19 - 20 - 21 - 22 - 23 - 24 - 25 -
(2π) $(n-1)$ $f(x)$ $(n+1)$ s^{-1} (x_0) $U(1)$ $\int_0^t$ $\int_0^\infty$ $\sum_{i=1}^n$ $g_{\mu\nu}$ $\frac{d}{dt}$ h^{-1} $2n+1$ $\int_{-\infty}^\infty$ $SU(2)$ $(k+1)$ $(k-1)$ g^{-1} e^+e^- (x_1) $(p-1)$ $\lim_{n\to\infty}$ (t') $(\mathbb{R}^n)$
$f(x)$ $f(z)$ $x(t)$ $(\mathbb{R}^n)$ $(n-1)$ $(n+1)$ $V(G)$ $u(t)$ a_{ij} $f(t)$ $2n+1$ $g(x)$ $u(x)$ $f(u)$ $F(x)$ u_{xx} $u(0)$ $p(x)$ $(\log n)$ $G=(V,E)$ $O(1)$ $|\nabla u|$ $y(t)$ (x_n) x_{n+1}
Complexity 4
arXMLiv
zbMATH
1 - 2 - 3 - 4 - 5 - 6 - 7 - 8 - 9 - 10 - 11 - 12 - 13 - 14 - 15 - 16 - 17 - 18 - 19 - 20 - 21 - 22 - 23 - 24 - 25 -
$(2\pi)^3$ $\int_{-\infty}^\infty$ $\sum_{i=1}^n$ $\frac{d}{dt}$ $(-1)^n$ $(2\pi)^4$ $(2\pi)^2$ $(t-t')$ $(2n+1)$ $(-1)^k$ $\sum_{n=1}^\infty$ $\sum_{n=0}^\infty$ $\sum_{i=1}^N$ $\frac{\partial}{\partial t}$ $\int_{\mathbb{R}^d}$ $|\nabla u|^2$ $1+o(1)$ $SU(2)_L$ kms^{-1} $\frac{1}{2\pi}$ $(t-t_0)$ $\sum_{j=1}^n$ $E(B-V)$ $\int_0^{2\pi}$ $(-1)^j$
$(n\log n)$ $O(\log n)$ $O(n^2)$ $(t,x(t))$ $f(0)=0$ $|f(z)|$ $(x(t))$ $u(x,0)=u_0(x)$ $(2n+1)$ $|\nabla u|^2$ $|V(G)|$ $|f(x)|$ $(t,u(t))$ $n^{-1/2}$ $1-|z|^2$ $(-1)^n$ $1+o(1)$ (a_{ij}) $O(n^3)$ $L^2(\mathbb{R}^n)$ $|\nabla u|^{p-2}$ dx/dt $v\in V(G)$ $x(0)=x_0$ $L^p(\mathbb{R}^n)$

Figure 4: The top-25 most frequent expressions in arXiv (left) and zbMATH (right) for complexities 1-4. A line between both sets indicates a matching set. Bold lines indicate that the matches share a similar rank (distance of 0 or 1).

C3		C4		C5		C6		C7	
114.84	$(n!)$	129.44	$i,j=1,\ldots,n$	119.21	$\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$	110.83	$(1+\lvert z\rvert^2)^\alpha$	98.72	$\mathrm{div}\left(\lvert\nabla u\rvert^{p-2}\,\nabla u\right)$
108.85	ϕ^{-1}	108.52	x_{ij}	112.55	$\lvert f(z)\rvert^p$	105.69	$f\left(re^{i\theta}\right)$	–	
100.19	z^{n-1}	108.50	$\dot{x}=A(t)x$	110.52	$\left(1+\lvert x\rvert^2\right)$	94.14	$f(z)=z+\sum_{n=2}^\infty a_n z^n$	–	
100.06	(c_n)	106.66	$\lvert x-x_0\rvert$	109.19	$\lvert f(x)\rvert^p$	92.33	$\left(\lvert\nabla u\rvert^{p-2}\,\nabla u\right)$	–	
100.05	$B(G)$	105.52	S^{2n+1}	106.22	$\lvert\nabla u\rvert^2 dx$	87.27	$(\log n/\log\log n)$	–	
99.87	$\log_2 n$	104.91	$L^2(\mathbb{R}^2)$	102.86	$n(n-1)/2$	78.54	$O(n\log^2 n)$	–	
99.65	$\xi(x)$	103.70	$\dot{x}=Ax+Bu$	101.40	$O(n^{-1})$	–		–	

Table 2: Top $s(t,D)$ scores, where D is the set of all zbMATH documents with a minimum document frequency of 200, maximum document frequency of 500k, and a minimum complexity of 3.

between $f(x)$ rather than a *function application*. Investigating the most frequently used terms in zbMATH in Table 4 reveals that u is most likely considered to be a function in the dataset: $u(t)$ (rank 8), $u(x)$ (rank 13), u_{xx} (rank 16), $u(0)$ (rank 17), $|\nabla u|$ (rank 22). Manual investigations of extended lists reveal even more hits: $u_0(x)$ (rank 30), $-\Delta u$ (rank 32), and $u(x,t)$ (rank 33). Since all eight terms are among the most frequent 35 entries in zbMATH, it implies that u can most likely be considered to imply a function in zbMATH. Of course, this does not imply that u must always be a function in zbMATH (see $f(u)$ on rank 14 in C3), but this allows us to exploit probabilities for improving MathIR performance. For instance, if not stated otherwise, u could be interpreted as a function by default, which could help increase the precision of the aforementioned tools.

Figure 4 also demonstrates that our two datasets diverge for increasing complexities. Hence, we can assume that frequencies of less complex formulae are more topic-independent. Conversely, the more complex a math formula is, the more context-specific it is. In the following, we will further investigate this assumption by applying TF-IDF rankings on the distributions.

4 RELEVANCE RANKING FOR FORMULAE

Zipf's law encourages the idea of scoring the relevance of words according to their number of occurrences in the corpus and in the documents. The family of BM25 ranking functions based on TF-IDF scores are still widely used in several retrieval systems [3, 34]. Since we demonstrated that mathematical formulae (and their subexpressions) obey Zipf's law in large scientific corpora, it appears intuitive

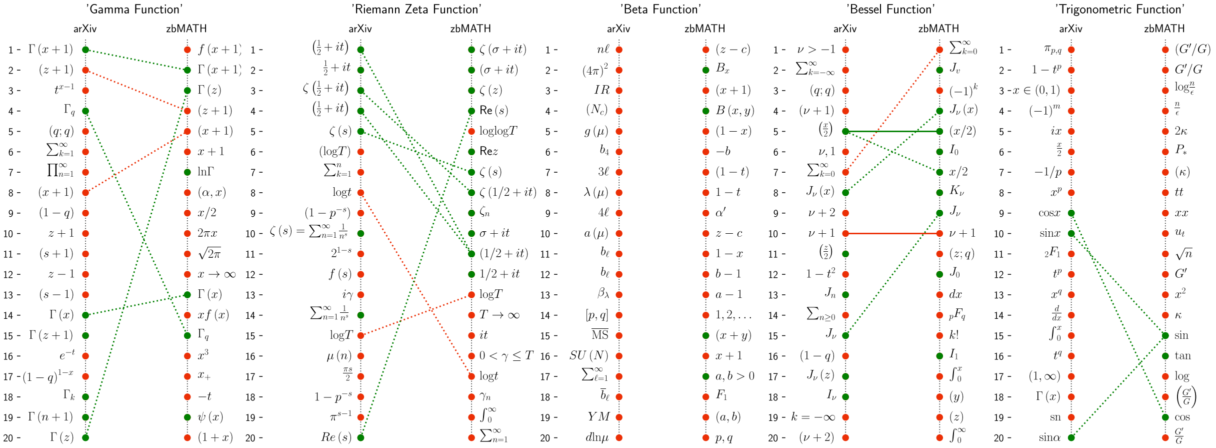


Figure 5: Top-20 ranked expressions retrieved from a topic-specific subset of documents D_q. The search query q is given above the plots. Retrieved formulae are annotated by a domain expert with green dots for relevant and red dots for non-relevant hits. A line is drawn if a hit appears in both result sets. The line is colored in green when the hit was marked as relevant.

to also use TF-IDF rankings, such as a variant of BM25, to calculate their relevance. In its original form [34], *Okapi BM25* was calculated as follows

$$\mathrm{bm25}(t,d) := \frac{(k+1)\,\mathrm{IDF}(t)\,\mathrm{TF}(t,d)}{\mathrm{TF}(t,d) + k\left(1 - b + \frac{b|d|}{\mathrm{AVG}_{\mathrm{DL}}}\right)}, \tag{7}$$

where $\mathrm{TF}(t,d)$ is the term frequency of t in the document d, $|d|$ the length of the document d (in our case, the number of subexpressions), $\mathrm{AVG}_{\mathrm{DL}}$ the average length of the documents in the corpus (see Table 1), and $\mathrm{IDF}(t)$ is the inverse document frequency of t, defined as

$$\mathrm{IDF}(t) := \log \frac{N - n(t) + \frac{1}{2}}{n(t) + \frac{1}{2}}, \tag{8}$$

where N is the number of documents in the corpus and $n(t)$ the number of documents which contain the term t. By adding $\frac{1}{2}$, we avoid $\log 0$ and division by 0. The parameters k and b are free, with b controlling the influence of the normalized document length and k controlling the influence of the term frequency on the final score. For our experiments, we chose the standard value $k = 1.2$ and a high impact factor of the normalized document length via $b = 0.95$.

As a result of our subexpression extraction algorithm, we generated a bias towards low complexities. Moreover, longer documents generally consist of more complex expressions. As demonstrated in Section 2.1, a document that only consists of the single expression $P_n^{(\alpha,\beta)}(x)$, i.e., the document had a length of one, would generate eight subexpressions, i.e., it results in a document length of eight. Thus, we modify the BM25 score in Equation (7) to emphasize higher complexities and longer documents. First, the average document length is divided by the average complexity $\mathrm{AVG}_{\mathrm{C}}$ in the corpus that is used (see Table 1), and we calculate the reciprocal of the document length normalization to emphasize longer documents.

Moreover, in the scope of a single document, we want to emphasize expressions that do not appear frequently in this document, but are the most frequent among their level of complexity. Thus, less complex expressions are ranked more highly if the document overall is not very complex. To achieve this weighting, we normalize the term frequency of an expression t according to its complexity $c(t)$ and introduce an inverse term frequency according to all expressions in the document

	arXiv	zbMATH
Retrieved Doc.	40	200
Min. Hit Freq.	7	7
Min. DF	50	10
Max. DF	10k	10k

Table 3: Settings for the retrieval experiments.

$$\mathrm{ITF}(t,d) := \log \frac{|d| - \mathrm{TF}(t,d) + \frac{1}{2}}{\mathrm{TF}(t,d) + \frac{1}{2}}. \tag{9}$$

Finally, we define the score $\mathrm{s}(t,d)$ of a term t in a document d as

$$s(t,d) := \frac{(k+1)\,\mathrm{IDF}(t)\,\mathrm{ITF}(t,d)\,\mathrm{TF}(t,d)}{\max\limits_{t' \in d|_{c(t)}} \mathrm{TF}(t',d) + k\left(1 - b + \frac{b\,\mathrm{AVG}_{\mathrm{DL}}}{|d|\,\mathrm{AVG}_{\mathrm{C}}}\right)}. \tag{10}$$

The TF-IDF ranking functions and the introduced $\mathrm{s}(t,d)$ are used to retrieve relevant documents for a given search query. However, we want to retrieve relevant subexpressions over a set of documents. Thus, we define the score of a formula (mBM25) over a set of documents as the maximum score over all documents

$$\mathrm{mBM25}(t,D) := \max_{d \in D} \mathrm{s}(t,d), \tag{11}$$

where D is a set of documents. We used *Apache Flink* [16] to count the expressions and process the calculations. Thus, our implemented system scales well for large corpora.

Table 2 shows the top-7 scored expressions, where D is the entire zbMATH dataset. The retrieved expressions can be considered as meaningful and real-world examples of MOIs, since most expressions are known for specific mathematical concepts, such as

Riemann Zeta Function

C1		C2		C3		C4		C5		C6		TF-IDF	mBM25
15,051	n	4,663	(s)	1,456	$\zeta(s)$	349	$(\frac{1}{2}+it)$	203	$\zeta(\frac{1}{2}+it)$	105	$\lvert\zeta(1/2+it)\rvert$	$\zeta(s)$	$\zeta\,(1/2+it)$
11,709	s	2,460	(x)	340	$\sigma+it$	232	$(1/2+it)$	166	$\zeta(1/2+it)$	88	$\left\lvert\zeta(\frac{1}{2}+it)\right\rvert$	$\zeta(1/2+it)$	$(1/2+it)$
9,768	x	2,163	(n)	310	$\sum_{n=1}^{\infty}$	195	$(\sigma+it)$	124	$\zeta(\sigma+it)$	81	$\lvert\zeta(\sigma+it)\rvert$	$(1/2+it)$	$(\frac{1}{2}+it)$
8,913	k	1,485	(t)	275	$(\log T)$	136	$\frac{1}{2}+it$	54	$\zeta(1+it)$	32	$\lvert\zeta(1+it)\rvert$	$\frac{1}{2}+it$	$\zeta\,(\frac{1}{2}+it)$
8,634	T	1,415	it	264	$1/2+it$	97	$s=\sigma+it$	44	$\zeta(2n+1)$	22	$\lvert\zeta(+it)\rvert$	$(\frac{1}{2}+it)$	$(\sigma+it)$

Eigenvalue

C1		C2		C3		C4		C5		C6		TF-IDF	mBM25
45,488	n	12,515	(x)	686	$-\Delta u$	218	$\lvert\nabla u\rvert^{p-2}$	139	$\lvert\nabla u\rvert^{p-2}\,\nabla u$	137	$\left(\lvert\nabla u\rvert^{p-2}\,\nabla u\right)$	$Ax=\lambda Bx$	$-\operatorname{div}\left(\lvert\nabla u\rvert^{p-2}\,\nabla u\right)$
43,090	x	6,598	(t)	555	$(n-1)$	218	$-\Delta_p u$	68	$-d^2/dx^2$	35	$-(py')'$	$-\Delta p$	$\operatorname{div}\left(\lvert\nabla u\rvert^{p-2}\,\nabla u\right)$
37,434	λ	4,377	λ_1	521	$\lvert\nabla u\rvert$	133	$W_0^{1,p}(\Omega)$	51	$A=(a_{ij})$	26	$(\lvert u'\rvert^{p-2}\,u')$	$P(\lambda)$	$p=\frac{N+2}{N-2}$
35,302	u	2,787	(Ω)	512	a_{ij}	127	$\lvert\nabla u\rvert^2$	46	$-\frac{d^2}{dx^2}$	18	$(\phi_p(u'))'$	λ_{k+1}	$\left(\phi_p\,(u')\right)'$
22,460	t	2,725	$\mathbb{R}^n$	495	$u(x)$	97	(a_{ij})	45	$u\in W_0^{1,p}(\Omega)$	18	$\int_\Omega \lvert\nabla u\rvert^2\,dx$	$\lambda_1>0$	$\lambda\in(0,\lambda^*)$

Table 4: The top-5 frequent mathematical expressions in the result set of zbMATH for the search queries 'Riemann Zeta Function' (top) and 'Eigenvalue' (bottom) grouped by their complexities (left) and the hits reordered according to their relevance scores (right). The TF-IDF score was calculated with normalized term frequencies.

$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, which refers to the Galois group of $\overline{\mathbb{Q}}$ over $\mathbb{Q}$, or $L^2(\mathbb{R}^2)$, which refers to the L^2-space (also known as *Lebesgue space*) over $\mathbb{R}^2$. However, a more topic-specific retrieval algorithm is desirable. To achieve this goal, we (i) retrieved a topic-specific subset of documents $D_q \subset D$ for a given textual search query q, and (ii) calculated the scores of all expressions in the retrieved documents. To generate D_q, we indexed the text sources of the documents from arXiv and zbMATH via elasticsearch (ES)[13] and performed the pre-processing steps: filtering stop words, stemming, and ASCII-folding[14]. Table 3 summarizes the settings we used to retrieve MOIs from a topic-specific subset of documents D_q. We also set a minimum hit frequency according to the number of retrieved documents an expression appears in. This requirement filters out uncommon notations.

Figure 5 shows the results for five search queries. We asked a domain expert from the National Institute of Standards and Technology (NIST) to annotate the results as related (shown as green dots in Figure 5) or non-related (red dots). We found that the results range from good performances (e.g., for the Riemann zeta function) to bad performances (e.g., beta function). For instance, the results for the Riemann zeta function are surprisingly accurate, since we could discover that parts of Riemann's hypothesis[15] were ranked highly throughout the results (e.g., $\zeta(\frac{1}{2}+it)$). On the other hand, for the beta function, we retrieved only a few related hits, of which only one had a strong connection to the beta function $B(x,y)$. We observed that the results were quite sensitive to the chosen settings (see Table 3). For instance, according to the beta function, the minimum hit frequency has a strong effect on the results, since many expressions are shared among multiple documents. For arXiv, the expressions $B(\alpha,\beta)$ and $B(x,y)$ only appear in one document of the retrieved 40. However, decreasing the minimum hit frequency would increase noise in the results.

Even though we asked a domain expert to annotate the results as relevant or not, there is still plenty of room for discussion. For instance, $(x+y)$ (rank 15 in zbMATH, 'Beta Function') is the argument of the gamma function $\Gamma(x+y)$ that appears in the definition of the beta function [8, (5.12.1)] $B(x,y) := \Gamma(x)\Gamma(y)/\Gamma(x+y)$. However, this relation is weak at best, and thus might be considered as not related. Other examples are $\mathrm{Re}z$ and $\mathrm{Re}(s)$, which play a crucial role in the scenario of the Riemann hypothesis (all non-trivial zeroes have $\mathrm{Re}(s) = \frac{1}{2}$). Again, this connection is not obvious, and these expressions are often used in multiple scenarios. Thus, the domain expert did not mark the expressions as being related.

Considering the differences in the documents, it is promising to have observed a relatively high number of shared hits in the results. Further, we were able to retrieve some surprisingly good insights from the results, such as extracting the full definition of the Riemann zeta function [8, (25.2.1)] $\zeta(s) := \sum_{n=1}^{\infty}\frac{1}{n^s}$. Even though a high number of shared hits seem to substantiate the reliability of the system, there were several aspects that affected the outcome negatively, from the exact definition of the search queries to retrieve documents via ES, to the number of retrieved documents, the minimum hit frequency, and the parameters in mBM25.

5 APPLICATIONS

The presented results are beneficial for a variety of use-cases. In the following, we will demonstrate and discuss several of the applications that we propose.

Extension of zbMATH's Search Engine: Formula search engines are often counterintuitive when compared to textual search, since the user must know how the system operates to enter a search query properly (e.g., does the system supports LaTeX inputs?). Additionally, mathematical concepts can be difficult to capture using only mathematical expressions. Consider, for example, someone

[13] https://github.com/elastic/elasticsearch [Accessed Sep. 2019]. We used version 7.0.0

[14] This means that non-ASCII characters are replaced by their ASCII counterparts or will be ignored if no such counterpart exists.

[15] Riemann proposed that the real part of every non-trivial zero of the Riemann zeta function is 1/2. If this hypothesis is correct, all the non-trivial zeros lie on the critical line consisting of the complex numbers $1/2 + it$.

Auto-completion for '$E = m$'			Suggestions for '$E = \{m, c\}$'		
Sug. Expression	TF	DF	Sug. Expression	TF	DF
$E = mc^2$	558	376	$E = mc^2$	558	376
$E = m\cosh\theta$	23	23	$E = \gamma mc^2$	39	38
$E = mv_0$	7	7	$E = \gamma m_e c^2$	41	36
$E = m/\sqrt{1-\dot{q}^2}$	12	6	$E = m\cosh\theta$	23	23
$E = m/\sqrt{1-\beta^2}$	10	6	$E = -mc^2$	35	17
$E = mc^2\gamma$	6	6	$E = \sqrt{m^2c^4 + p^2c^2}$	10	8

Table 5: Suggestions to complete '$E = m$' and '$E = \{m, c\}$' (the right-hand side contains m and c) with term and document frequency based on the distributions of formulae in arXiv.

who wants to search for mathematical expressions that are related to eigenvalues. A textual search query would only retrieve entire documents that require further investigation to find related expressions. A mathematical search engine, on the other hand, is impractical since it is not clear what would be a fitting search query (e.g., $Av = \lambda v$?). Moreover, formula and textual search systems for scientific corpora are separated from each other. Thus, a textual search engine capable of retrieving mathematical formulae can be beneficial. Also, many search engines allow for narrowing down relevant hits by suggesting filters based on the retrieved results. This technique is known as faceted search. The zbMATH search engine also provides faceted search, e.g., by authors, or year. Adding facets for mathematical expressions allows users to narrow down the results more precisely to arrive at specific documents.

Our proposed system for extracting relevant expressions from scientific corpora via mBM25 scores can be used to search for formulae even with textual search queries, and to add more filters for faceted search implementations. Table 4 shows two examples of such an extension for zbMATH's search engine. Searching for 'Riemann Zeta Function' and 'Eigenvalue' retrieved 4,739 and 25,248 documents from zbMATH, respectively. Table 4 shows the most frequently used mathematical expressions in the set of retrieved documents. It also shows the reordered formulae according to a default TF-IDF score (with normalized term frequencies) and our proposed mBM25 score. The results can be used to add filters for faceted search, e.g., show only the documents which contain $u \in W_0^{1,p}(\Omega)$. Additionally, the search system now provides more intuitive textual inputs even for retrieving mathematical formulae. The retrieved formulae are also interesting by themselves, since they provide insightful information on the retrieved publications. As already explored with our custom document search system in Figure 5, the Riemann hypothesis is also prominent in these retrieved documents.

The differences between TF-IDF and mBM25 ranking illustrates the problem of an extensive evaluation of our system. From a broader perspective, the hit $Ax = \lambda Bx$ is highly correlated with the input query 'Eigenvalue'. On the other hand, the raw frequencies revealed a prominent role of $\mathrm{div}(|\nabla u|^{p-2}\,\nabla u)$. Therefore, the top results of the mBM25 ranking can also be considered as relevant.

Math Notation Analysis: A faceted search system allows us to analyze mathematical notations in more detail. For instance, we can retrieve documents from a specific time period. This allows one to study the evolution of mathematical notation over time [4], or for identifying trends in specific fields. Also, we can analyze standard notations for specific authors since it is often assumed that authors prefer a specific notation style which may vary from the standard notation in a field.

Math Recommendation Systems: The frequency distributions of formulae can be used to realize effective math recommendation tasks, such as type hinting or error-corrections. These approaches require long training on large datasets, but may still generate meaningless results, such as $G_i = \{(x, y) \in \mathbb{R}^n : x_i = x_i\}$ [42]. We propose a simpler system which takes advantage of our frequency distributions. We retrieve entries from our result database, which contain all unique expressions and their frequencies. We implemented a simple prototype that retrieves the entries via pattern matching. Table 5 shows two examples. The left side of the table shows suggested autocompleted expressions for the query '$E{=}m$'. The right side shows suggestions for '$E{=}$', where the right-hand side of the equation should contain m and c in any order. A combination using more advanced retrieval techniques, such as similarity measures based on symbol layout trees [7, 45], would enlarge the number of suggestions. This kind of autocomplete and error-correction type-hinting system would be beneficial for various use-cases, e.g., in educational software or for search engines as a pre-processing step of the input.

Plagiarism Detection Systems: As previously mentioned, plagiarism detection systems [28, 29, 39] would benefit from a system capable of distinguishing conventional from uncommon notations. The approaches described by Meuschke et al. [29] outperform existing approaches by considering frequency distributions of single identifiers (expressions of complexity one). Considering that single identifiers make up only 0.03% of all unique expressions in arXiv, we presume that better performance can be achieved by considering more complex expressions. The conferred string representation also provides a simple format to embed complex expressions in existing learning algorithms.

Expressions with high complexities that are shared among multiple documents may provide further hints to investigate potential plagiarisms. For instance, the most complex expression that was shared among three documents in arXiv was Equation (3). A complex expression being identical in multiple documents could indicate a higher likelihood of plagiarism. Further investigation revealed that similar expressions, e.g., with infinite sums, are frequently used among a larger set of documents. Thus, the expression seems to be a part of a standard notation that is commonly shared, rather than a good candidate for plagiarism detection. Resulting from manual investigations, we could identify the equation as part of a concept called *generalized Hardy-Littlewood inequality* and Equation (3) appears in the three documents [2, 5, 32]. All three documents shared one author in common. Thus, this case also demonstrates a correlation between complex mathematical notations and authorship.

Semantic Taggers and Extraction Systems: We previously mentioned that semantic extraction systems [23, 36, 37] and semantic math taggers [6, 43] have difficulties in extracting the essential components (MOIs) from complex expressions. Considering the definition of the Jacobi polynomial in Equation (1), it would be beneficial to extract the groups of tokens that belong together, such as $P_n^{(\alpha,\beta)}(x)$ or $\Gamma(\alpha + m + 1)$. With our proposed search engine

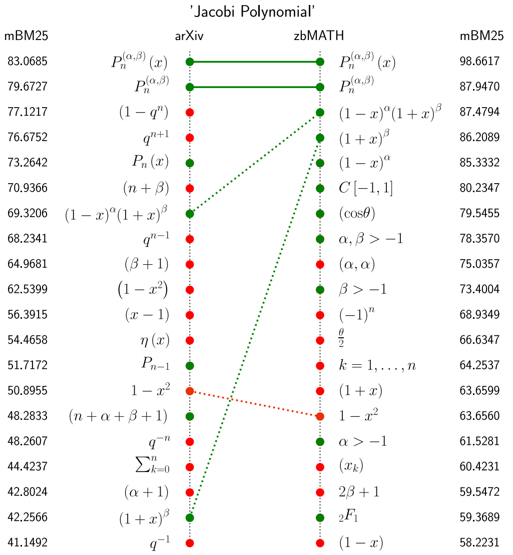


Figure 6: The top ranked expression for '*Jacobi polynomial*' in arXiv and zbMATH. For arXiv, 30 documents were retrieved with a minimum hit frequency of 7.

for retrieving MOIs, we are able to facilitate semantic extraction systems and semantic math taggers. Imagine such a system being capable of identifying the term 'Jacobi polynomial' from the textual context. Figure 6 shows the top relevant hits for the search query 'Jacobi polynomial' retrieved from zbMATH and arXiv. The results contain several relevant and related expressions, such as the constraints $\alpha, \beta > -1$ and the weight function for the Jacobi polynomial $(1-x)^\alpha(1+x)^\beta$, which are essential properties of this orthogonal polynomial. Based on these retrieved MOIs, the extraction systems can adjust its retrieved math elements to improve precision, and semantic taggers or a tokenizer could re-organize parse trees to more closely resemble expression trees.

6 CONCLUSION & FUTURE WORK

In this study we showed that analyzing the frequency distributions of mathematical expressions in large scientific datasets can provide useful insights for a variety of applications. We demonstrated the versatility of our results by implementing prototypes of a type-hinting system for math recommendations, an extension of zbMATH's search engine, and a mathematical retrieval system to search for topic-specific MOIs. Additionally, we discussed the potential impact and suitability in other applications, such as math search engines, plagiarism detection systems, and semantic extraction approaches. We are confident that this project lays a foundation for future research in the field of MathIR.

We plan on developing a web application which would provide easy access to our frequency distributions, the MOI search engine, and the type-hinting recommendation system. We hope that this will further expedite related future research projects. Moreover, we will use this web application for an online evaluation of our MOI retrieval system. Since the level of agreement among annotators will be predictably low, an evaluation by a large community is desired.

In this first study, we preserved the core structure of the MathML data which provided insightful information for the MathML community. However, this makes it difficult to properly merge formulae. In future studies, we will normalize the MathML data via MathMLCan [9]. In addition to this normalization, we will include wildcards for investigating distributions of formula patterns rather than exact expressions. This will allow us to study connections between math objects, e.g., between $\Gamma(z)$ and $\Gamma(x+1)$. This would further improve our recommendation system and would allow for the identification of regions for parameters and variables in complex expressions.

ACKNOWLEDGMENTS

Discovering Mathematical Objects of Interest was supported by the German Research Foundation (DFG grant GI-1259-1).

REFERENCES

[1] Akiko N. Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Inf. Process. Manage.* Vol. 39, Issue 1, 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3

[2] Gustavo Araujo and Daniel Pellegrino. 2014. On the constants of the Bohnenblust-Hille inequality and Hardy–Littlewood inequalities. In *Computing Research Repository (CoRR)* *https://arxiv.org/abs/1407.7120.*

[3] Jöran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *Int. J. on Digital Libraries* Vol. 17, Issue 4, 305–338. https://doi.org/10.1007/s00799-015-0156-0

[4] Florian Cajori. 1929. *A History of Mathematical Notations.* The Open Court Company, London, UK. Vol. 1 & 2.

[5] Jamilson R. Campos, Wasthenny Cavalcante, Vinícius V. Fávaro, Daniel Nu nez Alarcón, Daniel Pellegrino, and Diana M. Serrano-Rodríguez. 2015. Polynomial and multilinear Hardy–Littlewood inequalities: analytical and numerical approaches. In *Computing Research Repository (CoRR)* *https://arxiv.org/abs/1503.00618.*

[6] Pao-Yu Chien and Pu-Jen Cheng. 2015. Semantic Tagging of Mathematical Expressions. In *Proc. WWW'2015.* ACM, 195–204. https://doi.org/10.1145/2736277.2741108

[7] Kenny Davila and Richard Zanibbi. 2017. Layout and Semantics: Combining Representations for Mathematical Formula Search. In *Proc. ACM SIGIR.* ACM, 1165–1168. https://doi.org/10.1145/3077136.3080748

[8] DLMF 2019. *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.0.25 of 2019-12-15. http://dlmf.nist.gov F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[9] David Formánek, Martin Líška, Michal Růžička, and Petr Sojka. 2012. Normalization of Digital Mathematics Library Content. In *Proc. of OpenMath/ MathUI/ CICM-WiP (CEUR Workshop Proceedings).* 91–103. http://ceur-ws.org/Vol-921/wip-05.pdf

[10] Liangcai Gao, Zhuoren Jiang, Yue Yin, Ke Yuan, Zuoyu Yan, and Zhi Tang. 2017. Preliminary Exploration of Formula Embedding for Mathematical Information Retrieval: can mathematical formulae be embedded like a natural language? In *Computing Research Repository (CoRR)* *http://arxiv.org/abs/1707.05154.*

[11] Alex Gaudeul. 2007. Do Open Source Developers Respond to Competition?: The LaTeX Case Study. In *Review of Network Economics* https://doi.org/10.2202/1446-9022.1119

[12] Deyan Ginev. 2018. *arXMLiv:08.2018 dataset, an HTML5 conversion of arXiv.org.* https://sigmathling.kwarc.info/resources/arxmliv/ SIGMathLing – Special Interest Group on Math Linguistics.

[13] André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William I. Grosky, and Bela Gipp. 2019. Why Machines Cannot Learn Mathematics, Yet. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019* Vol. 2414. CEUR-WS.org, 130–137. http://ceur-ws.org/Vol-2414/paper14.pdf

[14] Christian Grün, Sebastian Gath, Alexander Holupirek, and Marc Scholl. 2009. XQuery Full Text Implementation in BaseX. In *Database and XML Technologies.* Springer Berlin, 114–128.

[15] Ferruccio Guidi and Claudio Sacerdoti Coen. 2016. A Survey on Retrieval of Mathematical Knowledge. *Mathematics in Computer Science* Vol. 10, Issue 4, 409–427. https://doi.org/10.1007/s11786-016-0274-0

[16] Fabian Hueske and Timo Walther. 2019. Apache Flink. In *Encyclopedia of Big Data Technologies.*, Sherif Sakr and Albert Y. Zomaya (Eds.). Springer. https://doi.org/10.1007/978-3-319-63962-8_303-1

[17] Shahab Kamali and Frank Wm. Tompa. 2010. A new mathematics retrieval system. In *Proc. ACM CIKM.* ACM, 1413–1416. https://doi.org/10.1145/1871437.1871635

[18] Shahab Kamali and Frank Wm. Tompa. 2013. Retrieving documents with mathematical content. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013.* ACM, 353–362. https://doi.org/10.1145/2484028.2484083

[19] Andrea Kohlhase. 2018. Factors for Reading Mathematical Expressions. In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", LWDA 2018, Mannheim, Germany, August 22-24, 2018. (CEUR Workshop Proceedings)*, Vol. 2191. CEUR-WS.org, 195–202. http://ceur-ws.org/Vol-2191/paper24.pdf

[20] Andrea Kohlhase, Michael Kohlhase, and Michael Fürsich. 2017. Visual Structure in Mathematical Expressions. In *Intelligent Computer Mathematics - 10th International Conference, CICM 2017, Edinburgh, UK, July 17-21, 2017, Proceedings (Lecture Notes in Computer Science)*, Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke (Eds.). Springer, 208–223. https://doi.org/10.1007/978-3-319-62075-6_15

[21] Andrea Kohlhase, Michael Kohlhase, and Taweechai Ouypornkochagorn. 2018. Discourse Phenomena in Mathematical Documents. In *Intelligent Computer Mathematics - 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, Proceedings (Lecture Notes in Computer Science)*, Vol. 11006. Springer, 147–163. https://doi.org/10.1007/978-3-319-96812-4_14

[22] Michael Kohlhase, Bogdan A. Matican, and Corneliu-Claudiu Prodescu. 2012. MathWebSearch 0.5: Scaling an Open Formula Search Engine. In *Intelligent Computer Mathematics - 11th International Conference, AISC 2012, 19th Symposium, Calculemus 2012, 5th International Workshop, DML 2012, 11th International Conference, MKM 2012, Systems and Projects, Held as Part of CICM 2012, Bremen, Germany, July 8-13, 2012. Proceedings.* Springer Berlin Heidelberg, 342–357. https://doi.org/10.1007/978-3-642-31374-5_23

[23] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. 2017. Utilizing dependency relationships between math expressions in math IR. *Information Retrieval Journal* Vol. 20, Issue 2, 132–167. https://doi.org/10.1007/s10791-017-9296-8

[24] Giovanni Yoko Kristianto, Goran Topic, Florence Ho, and Akiko Aizawa. 2014. The MCAT Math Retrieval System for NTCIR-11 Math Track. In *Proc. 11th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences,*, Noriko Kando, Hideo Joho, and Kazuaki Kishida (Eds.). National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/06-NTCIR11-MATH-KristiantoGY.pdf

[25] Kriste Krstovski and David M. Blei. 2018. Equation Embeddings. In *Computing Research Repository (CoRR) http://arxiv.org/abs/1803.09123.*

[26] Aldo Lipani, Linda Andersson, Florina Piroi, Mihai Lupu, and Allan Hanbury. 2014. TUW-IMP at the NTCIR-11 Math-2. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, December 9-12, 2014.* National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/09-NTCIR11-MATH-LipaniA.pdf

[27] Ashish Lohia, Kirti Sinha, Soujanya Vadapalli, and Kamalakar Karlapalem. 2005. An Architecture for Searching and Indexing Latex Equations in Scientific Literature. In *Proc. COMAD.* Computer Society of India, 122–130. http://comad2005.persistent.co.in/COMAD2005Proc/pages122-130.pdf

[28] Norman Meuschke, Moritz Schubotz, Felix Hamborg, Tomás Skopal, and Bela Gipp. 2017. Analyzing Mathematical Content to Detect Academic Plagiarism. In *Proc. ACM CIKM.* ACM, 2211–2214. https://doi.org/10.1145/3132847.3133144

[29] Norman Meuschke, Vincent Stange, Moritz Schubotz, Michael Kramer, and Bela Gipp. 2019. Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL).* 120–129. https://doi.org/10.1109/JCDL.2019.00026

[30] Bruce R. Miller. 2019. `LaTeXML` A LaTeX to XML/HTML/MathML Converter. http://dlmf.nist.gov/LaTeXML/. http://dlmf.nist.gov/LaTeXML/ Accessed: 2019-09-01.

[31] Shunsuke Ohashi, Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. 2016. Efficient Algorithm for Math Formula Semantic Search. *IEICE Transactions* Vol. E99.D, Issue 4, 979–988. https://doi.org/10.1587/transinf.2015DAP0023

[32] Daniel Pellegrino. 2015. A short communication on the constants of the multilinear Hardy–Littlewood inequality. In *Computing Research Repository (CoRR) https://arxiv.org/abs/1510.00367.*

[33] Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* Vol. 21, Issue 5 (March 2014), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

[34] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* Vol. 3, Issue 4, 333–389. https://doi.org/10.1561/1500000019

[35] Moritz Schubotz, André Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard S. Cohl, and Bela Gipp. 2018. Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*, Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, and Vivien Petras (Eds.). ACM, 233–242. https://doi.org/10.1145/3197026.3197058

[36] Moritz Schubotz, Alexey Grigorev, Marcus Leich, Howard S. Cohl, Norman Meuschke, Bela Gipp, Abdou S. Youssef, and Volker Markl. 2016. Semantification of Identifiers in Mathematics for Better Math Information Retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16).* ACM, 135–144. https://doi.org/10.1145/2911451.2911503 Full Paper.

[37] Moritz Schubotz, Leonard Krämer, Norman Meuschke, Felix Hamborg, and Bela Gipp. 2017. Evaluating and Improving the Extraction of Mathematical Identifier Definitions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings (Lecture Notes in Computer Science)*, Gareth J. F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro (Eds.), Vol. 10456. Springer, 82–94. https://doi.org/10.1007/978-3-319-65813-1_7

[38] Moritz Schubotz and Olaf Teschke. 2019. Four decades of TeX at zbMATH. *Newsletter of the European Mathematical Society (EMS)* 6, 50–52. https://doi.org/10.4171/NEWS/112/15

[39] Moritz Schubotz, Olaf Teschke, Vincent Stange, Norman Meuschke, and Bela Gipp. 2019. Forms of Plagiarism in Digital Mathematical Libraries. In *Intelligent Computer Mathematics - 12th International Conference, CICM 2019, Prague, Czech Republic, July 8-12, 2019, Proceedings (Lecture Notes in Computer Science)*, Vol. 11617. Springer, 258–274. https://doi.org/10.1007/978-3-030-23250-4_18

[40] Glenn Gordon Smith and David Ferguson. 2004. Diagrams and math notation in e-learning: growing pains of a new generation. *International Journal of Mathematical Education in Science and Technology* Vol. 35, 681–695. Issue 5. https://doi.org/10.1080/0020739042000232583

[41] Leonard Wörteler, Michael Grossniklaus, Christian Grün, and Marc Scholl. 2015. Function inlining in XQuery 3.0 optimization. In *Proc. 15th DBLP.* ACM, 45–48. https://doi.org/10.1145/2815072.2815079

[42] Michihiro Yasunaga and John Lafferty. 2019. TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts. In *Computing Research Repository (CoRR) http://arxiv.org/abs/1902.06034.*

[43] Abdou Youssef. 2017. Part-of-Math Tagging and Applications. In *Intelligent Computer Mathematics*, Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke (Eds.). Springer International Publishing, Cham, 356–374.

[44] Abdou Youssef and Bruce R. Miller. 2019. Explorations into the Use of Word Embedding in Math Search and Math Semantics. In *Intelligent Computer Mathematics - 12th International Conference, CICM 2019, Prague, Czech Republic, July 8-12, 2019, Proceedings (Lecture Notes in Computer Science)*, Vol. 11617. Springer, 291–305. https://doi.org/10.1007/978-3-030-23250-4_20

[45] Richard Zanibbi, Kenny Davila, Andrew Kane, and Frank Wm. Tompa. 2016. Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16).* ACM, New York, NY, USA, 145–154. https://doi.org/10.1145/2911451.2911512

Evaluating Unicast and MBSFN in Public Safety Networks

Chunmei Liu*, Chen Shen*, Jack Chuang*, Richard A. Rouil*, and Hyeong-Ah Choi†
*Wireless Networks Division, National Institute of Standards and Technology, USA
†Department of Computer Science, George Washington University, USA
Email: *{chunmei.liu, chen.shen, jack.chaung, richard.rouil}@nist.gov, †{hchoi}@gwu.edu

***Abstract*—Public safety incidents typically involve significant amount of group traffic and have a stringent requirement on connection reliability. Hence multicast could potentially improve network and user performance significantly. Towards this goal, in this paper we investigate Long-Term Evolution (LTE) Multicast Broadcast Single Frequency Network (MBSFN) from the perspectives of throughput, resource efficiency, sources of MBSFN gain, and outage. Firstly, we derive a realistic MBSFN Signal-to-Interference-plus-Noise Ratio (SINR) analytical model, with multiple antennas, multipath channel, and equalizer considered. Secondly, we develop an MBSFN system level simulator as well as a simulation platform to compare unicast and MBSFN performance. Thirdly, we perform comprehensive simulations on the metrics under study. Through simulations, we discover that while unicast may achieve higher information bits/symbol, MBSFN provides higher throughput, and MBSFN throughput increases with MBSFN area size. We then quantify the MBSFN SINR improvement due to diversity combining and interference reduction, respectively. Furthermore, we show that compared with unicast, MBSFN improves outage probability significantly, which is essential for public safety incidents. Finally, we show that all the above MBSFN performance improvements apply to different MBSFN area sizes, even when the MBSFN size is one cell.**

***Index Terms*—Public safety, LTE, MBSFN, SINR analytical model, outage probability, resource efficiency.**

I. Introduction

Public safety mission is essential to protect citizens' lives and properties, and effective communications among first responders during public safety incidents is crucial. Compared with commercial traffic, public safety incidents typically involve significant amount of group traffic among first responders [1], including traffic intensive applications, such as mission critical video. Using traditional point-to-point unicast transmission at physical (PHY) layer to serve this type of traffic would require significant amount of spectrum, and sometimes lead to severe network congestion. With the above in mind, using multicast to serve public safety traffic has been put on the table, and Multicast Broadcast Single Frequency Network (MBSFN) in Long-Term Evolution (LTE) is one candidate due to its multicast nature and potential Signal-to-Interference-plus-Noise Ratio (SINR) improvement especially at cell edges [2] [3].

From a technology perspective, while MBSFN was developed in LTE and uses Single Frequency Network (SFN), SFN itself was not new and had been investigated decades ago for classical broadcast technologies, such as Digital Audio Broadcasting (DAB) and Digital Video Broadcasting (DVB) [4] [5]. These early works demonstrated that SFN could improve coverage and spectrum usage.

While both LTE MBSFN and classical broadcasting technologies use SFN, LTE MBSFN operates under LTE architecture and follows LTE protocol stacks. System performance is hence different, which triggered studies of LTE MBSFN in recent years from various perspectives. In [6], the authors first derived an analytical expression for SINR at a given point in a cell, then evaluated MBSFN performance under different Modulation and Coding Schemes (MCSs), the length of cyclic prefix, and the impact of shadowing. In [7], the authors extended the work in [6] to femtocell and compared single cell transmissions with multicell transmissions. None of the above work considered multiple antennas at the transmitter side or the receiver side. These multiple antennas allow multiple-input and multiple-output (MIMO) in unicast, which is one major technology that made LTE successful. In addition, the performance between MBSFN and unicast are compared in [8], where the authors used average MCS for all user equipments (UEs) for performance calculation. In [9], the authors proposed an optimal UE grouping algorithm in MBSFN, where the UEs were associated with the same MCS in unicast and multicast transmissions. As shown in our previous work [10], this is not always the case. MCSs could be different for different UEs, and MCSs used for unicast and multicast could be different for the same UE as well.

In this paper we investigate LTE MBSFN from the perspectives of throughput, resource efficiency, MBSFN gain, and outage. Unlike commercial broadcasting in a large deployment area, public safety incidents could vary in size and happen within a small area. Therefore we also consider small MBSFN areas. Specifically, we first derive an SINR analytical model for MBSFN. Different from others' work, multiple antennas, multipath channel, and equalizer are considered. The resulting model is hence more realistic and accurate. Next, by closely following the 3rd Generation Partnership Project (3GPP) standards, we implement the MBSFN SINR model and develop an MBSFN system level simulator as well as a simulation platform to compare unicast and MBSFN performance. Different from others' work, here SINR values in unicast and MBSFN are calculated separately based on their unique models. The high fidelity MBSFN Block Error Rate (BLER) curves and Channel Quality Indicator (CQI) switch points from our previous work [10] are also employed for PHY abstraction. Then, through comprehensive simulations, we show that although unicast may have higher resource efficiency in terms of information bits/symbol, MBSFN would provide higher throughput since it uses the full bandwidth in each transmission that serves all first responders, while in unicast resources are shared among first responders. Simulation

Liu, Chunmei; Shen, Chen; Chuang, Jack; Rouil, Richard A.; Choi, Hyeong-Ah. "Evaluating Unicast and MBSFN in Public Safety Networks."
Paper presented at IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2020, London, UK. August 31, 2020 - September 03, 2020.

results also confirmed that larger MBSFN area size would increase MBSFN throughput due to higher MBSFN diversity combining gain and less interference. We further quantify MBSFN SINR improvement from diversity combining and interference reduction, respectively. The results show that while at cell edge they are comparable, at cell center the gain from interference reduction dominates and is significantly higher. However, the large SINR gain at cell center would not improve performance significantly due to MCS cap and lack of MIMO support for MBSFN. Finally, simulation results show that compared with unicast, MBSFN improves outage probability significantly, which is essential for public safety incidents. The above MBSFN performance improvements hold even when the MBSFN area size is one cell, which applies to the case when the public safety incident area is small.

The rest of the paper is organized as follows. In Section II we briefly describe unicast and MBSFN details as specified in 3GPP that are relevant to our analysis. In Section III we derive the SINR analytical model for MBSFN. In Section IV we describe our simulation design, together with simulation results and analysis. In Section V we summarizes our findings.

II. Review of Unicast and MBSFN in LTE

In this section we review unicast and MBSFN as specified by 3GPP, with emphasis on factors that have direct impacts on our analysis. The public safety broadband network Band 14 (B14) 10 MHz bandwidth Frequency Division Duplexing (FDD) is used as an example [11], with a focus on downlink.

A. LTE Overview

3GPP specifies Orthogonal Frequency-Division Multiple Access (OFDMA) for LTE downlink, and Cyclic Prefix (CP) is used to avoid Inter-Symbol Interference (ISI) [12]. Both normal CP and extended CP are specified. While the normal CP is used in typical LTE deployments to achieve high data rate, the extended CP is used in special cases such as in very large cells and MBSFN. In this paper, normal CP is used in unicast analysis, and extended CP is used for MBSFN.

For FDD, 3GPP defines frame structure type 1 [13], where one radio frame is 10 ms in duration and divided into 10 subframes. The smallest time-frequency unit for downlink transmission is a resource element (RE), which consists of one Orthogonal Frequency Division Multiplex (OFDM) subcarrier for a duration of one OFDM symbol. Transmissions can be scheduled by Resource Blocks (RBs) [14], and data is carried in Transport Blocks (TBs), which are passed from the Media Access Control (MAC) layer to the PHY layer once per Transmission Time Interval (TTI) which is 1 ms.

UEs report channel quality back to eNodeB using different sets of CQI indices for different MCS sets the network may deploy. In this paper, we consider the MCS set of Quadrature Phase Shift Keying (QPSK), 16 Quadrature Amplitude Modulation (QAM), and 64QAM. The corresponding CQI indices and their interpretations are given in Table 7.2.3-1 in [13].

B. LTE Unicast Review

For unicast, RB is the basic unit when allocating resources. To improve data rate and robustness, LTE unicast utilizes MIMO technology such as transmit diversity, spatial multiplexing, and beamforming. Accordingly, 3GPP defines 10 transmission modes (TMs) and could support up to 8 layer transmissions using TM9 [14], which is the transmission mode used in this paper. In addition, since Hybrid Automatic Repeat Request (HARQ) is specified for unicast transmissions, in this paper the target BLER is selected to be 10 %.

C. LTE MBSFN Review

In MBSFN, data is transmitted from multiple cells to the destination UE, as shown in Figure 1. All cells involved are tightly synchronized and transmit the same content over the same subcarriers using the same waveform. These cells form a so-called MBSFN area. To avoid ISI, the transmissions from different cells are targeted to arrive at the UE within CP at the start of the symbol. Hence 3GPP specifies the extended CP for MBSFN. From the UE perspective, the UE could treat all the transmissions from MBSFN area in the same way as multipath components of a single cell transmission, and the SINR could improve especially at cell edge. Note that the SINR improvement comes from two folds. One is the diversity combining gain from multiple signal sources. Another is the reduction of inter-cell interference - the transmissions from additional MBSFN cells are now turned into constructive signals instead of inter-cell interference. In Section IV-C, we will quantify the SINR improvement from these two folds.

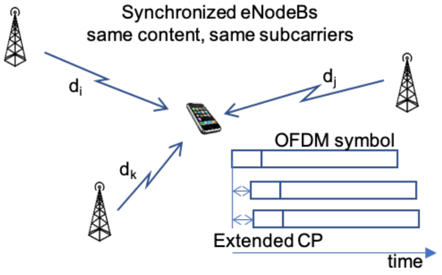


Fig. 1. MBSFN Transmission [15]

The composite channel from multiple cells in MBSFN requires a separate channel estimation from that performed from a single cell. 3GPP hence defines MBSFN subframe that carries MBSFN reference signals. MBSFN data is transmitted in MBSFN subframes only while unicast data is transmitted in non-MBSFN subframes, and MBSFN subframes and non-MBSFN subframes are interleaved in time. In addition, out of the 10 subframes within one radio frame, subframe 0, 4, 5, and 9 carry control information that is essential for network operation, such as paging occasions [14]. These four subframes are hence reserved for unicast transmissions and cannot be configured as MBSFN subframes. Therefore, there are at most six subframes available for MBSFN transmissions, or 60 % of the total resources.

3GPP also specifies that a single TB is generated per TTI for multicast channel (MCH) and uses all the MBSFN resources in that subframe [16] (e.g. 50 RBs for public safety band B14). In addition, no transmit diversity scheme is specified, and MCH is mapped on a single layer spatial multiplexing. Hence MBSFN could not take advantage of MIMO technologies. Furthermore, given its multicast nature, 3GPP specifies no radio link control retransmissions and no HARQ for MBSFN. Hence, in order to deliver acceptable service to upper layers, lower target BLER is typically used for MBSFN. In this paper we use 1 %.

III. MBSFN Analytical Model

In this section we derive an analytical model for MBSFN SINR, considering multiple antennas, multipath channel, and equalizer. We then convert it into Additive White Gaussian Noise (AWGN) equivalent SINR, which will later be fed into system level simulations.

A. *MBSFN Networks*

Consider a regular hexagonal network with three sectors per cell, as illustrated in Figure 2, where the numbers are cell identifiers (IDs). Let $\mathcal{I}$ denote the set of cells under consideration, and N denote the total number of cells in set $\mathcal{I}$. Index these N cells such that the first N_{M} cells are cells that participate in MBSFN transmissions within the MBSFN area (cell 1 to 21 in Figure 2), and denote them by set $\mathcal{I}_{\text{M}} = \{1, 2, \cdots, N_{\text{M}}\}$. Consequently, the cells with indices among $N_{\text{M}}+1, \cdots, N$ are the cells that do not participate in MBSFN transmissions (the cells that are not assigned an ID in Figure 2). Denote them by set $\mathcal{I}_{\text{L}} = \{N_{\text{M}}+1, N_{\text{M}}+2, \cdots, N\}$, where its size $N_{\text{L}} = N - N_{\text{M}}$. All cells have the same number of transmit antennas N_{Tx}, and all UEs have the same number of receiver antennas N_{Rx}.

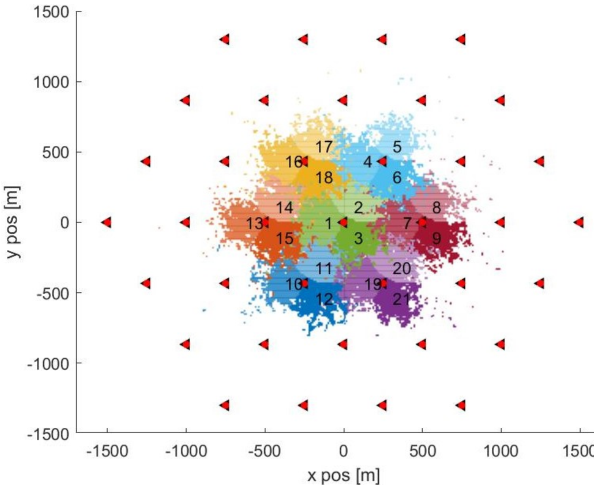


Fig. 2. Network Layout with Cell ID

Consider a UE within the MBSFN area. As shown in Figure 1, let d_i denote the distance between MBSFN cell $i \in \mathcal{I}_{\text{M}}$ and the UE, while MBSFN cell $k \in \mathcal{I}_{\text{M}}$ be the cell that is closest to the UE. As mentioned previously, in MBSFN, all MBSFN cells are tightly synchronized and transmit the same content simultaneously at the same subcarriers. Use the first signal received from cell k as the reference signal, and align to it all signals received after, including those from other MBSFN cells. Then the signals from cell i incur delay τ_i [6] [4]:

$$\tau_i = \frac{d_i - d_k}{s}, \tag{1}$$

where s is the speed of light. Note that given the relatively long CP length and signal frame - 16.67 μs for the extended CP length and 66.67 μs for the signal frame per 3GPP - the delays from different paths from the same cell i are not further differentiated, and are all approximated by the same τ_i.

The constructive and destructive portions of the signal from cell i, $i \neq k$ ($i \in \mathcal{I}_{\text{M}}$) can be captured by weight function ω_i and $1 - \omega_i$, respectively, where ω_i is as below [6] [4]:

$$\omega_i = \begin{cases} 1 & \text{when } 0 \leq \tau_i < T_{\text{CP}} \\ 1 - \dfrac{\tau_i - T_{\text{CP}}}{T_{\text{u}}} & \text{when } T_{\text{CP}} \leq \tau_i < T_{\text{CP}} + T_{\text{u}} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where T_{CP} is the extended CP length defined in 3GPP for MBSFN, and T_{u} is the length of the useful signal frame.

Note that the destructive portion leads to ISI for the next symbol.

B. *RE Level SINR*

As mentioned in Section II, 3GPP specifies that a single TB per TTI is used for MBSFN transmission, and this TB uses all resources in that subframe. Let N_{RE} denote the total number of REs within one MBSFN subframe, and index the REs by $c = 1, 2, \cdots, N_{\text{RE}}$. Then the $(N_{\text{Rx}} \times 1)$-element UE received signal vector at RE c, denoted by $\boldsymbol{y}^c$, can be expressed as

$$\begin{aligned} \boldsymbol{y}^c = & \sum_{i=\mathcal{I}_{\text{M}}} \sqrt{\omega_i P_i^c} \boldsymbol{H}^{(i,c)} \mathbf{1}_{N_{\text{Tx}}} x_m \\ & + \sum_{i=\mathcal{I}_{\text{M}}} \sqrt{(1-\omega_i) P_i^c} \boldsymbol{H}^{(i,c)} \mathbf{1}_{N_{\text{Tx}}} x_m^- \\ & + \sum_{l=\mathcal{I}_{\text{L}}} \sqrt{P_l^c} \boldsymbol{H}^{(l,c)} \boldsymbol{W}^{(l,c)} \boldsymbol{\chi}_l + \boldsymbol{n}^c, \end{aligned} \tag{3}$$

where x_m is the MBSFN transmit signal, x_m^- is the previous MBSFN transmit signal; $\boldsymbol{\chi}_l$ is the $(N_{\text{Tx}} \times 1)$-element transmit signal vector from cell l, $l \in \mathcal{I}_{\text{L}}$ (i.e. $l \notin \mathcal{I}_{\text{M}}$); P_i^c and P_l^c ($i \in \mathcal{I}_{\text{M}}, l \in \mathcal{I}_{\text{L}}$) are the signal powers from cell i and l at RE c after taking into account path loss and shadowing but without small-scale fading, respectively; $\boldsymbol{H}^{(i,c)}$ and $\boldsymbol{H}^{(l,c)}$, ($i \in \mathcal{I}_{\text{M}}$, $l \in \mathcal{I}_{\text{L}}$) are the $(N_{\text{Rx}} \times N_{\text{Tx}})$-element matrices representing the frequency domain channel gains from cell i and l to the UE at RE c, respectively; $\boldsymbol{W}^{(l,c)}$ is the $(N_{\text{Tx}} \times N_{\text{Tx}})$ channel precoding matrix used by cell l on RE c, $l \in \mathcal{I}_L$; $\mathbf{1}_{N_{\text{Tx}}}$ is $(N_{\text{Tx}} \times 1)$ column vector with all elements being 1; and $\boldsymbol{n}^c$ represents thermal noise with zero mean and variance σ_N^2.

The first term in Eq. (3) represents the constructive portion of the signal. The second term represents the ISI from the destructive portion of the previous transmit signal, where the channel experienced by this destructive portion is approximated by the channel experienced by the current signal. The third term represents interference from the non-MBSFN cells. And the last term represents the thermal noise. Note that all MBSFN cells and all transmit antennas at each MBSFN cell transmit the same signal. Hence x_m applies to all MBSFN cell $i = 1, \cdots, N_{\mathrm{M}}$ and each of their transmit antennas. x_m is hence essentially a scalar.

Define

$$\boldsymbol{q}^c = \left[\sqrt{\omega_1 P_1^c}\boldsymbol{H}^{(1,c)}, \cdots, \sqrt{w_{N_{\mathrm{M}}} P_{N_{\mathrm{M}}}^c}\boldsymbol{H}^{(N_{\mathrm{M}},c)}\right]\mathbf{1}_{(N_{\mathrm{Tx}})(N_{\mathrm{M}})}, \tag{4}$$

where $\mathbf{1}_{(N_{\mathrm{Tx}})(N_{\mathrm{M}})}$ is $(N_{\mathrm{Tx}} \cdot N_{\mathrm{M}} \times 1)$-element column vector with all elements being 1. Then:

$$\boldsymbol{y}^c = \boldsymbol{q}^c x_m + \sum_{i=1}^{N_{\mathrm{M}}} \sqrt{(1-\omega_i)P_i^c}\boldsymbol{H}^{(i,c)}\mathbf{1}_{N_{\mathrm{Tx}}} x_m^- + \sum_{l=N_{\mathrm{M}}+1}^{N} \sqrt{P_l^c}\boldsymbol{H}^{(l,c)}\boldsymbol{W}^{(l,c)}\boldsymbol{\chi}_l + \boldsymbol{n}^c. \tag{5}$$

Eq. (5) shows that the system could be viewed as a Single Input Multiple Output (SIMO) system, with signals from transmit antennas of all MBSFN cells being treated as the same transmit signal from different paths. We could hence apply zero-forcing receiver

$$\boldsymbol{f}^c = [(\boldsymbol{q}^c)^H \boldsymbol{q}^c]^{-1}(\boldsymbol{q}^c)^H, \tag{6}$$

where the superscript $(\cdot)^H$ represents conjugate transpose. Then the post-equalization received signal becomes

$$r^c = \boldsymbol{f}^c \boldsymbol{y}^c \tag{7}$$

$$= x_m + \sum_{i=1}^{N_{\mathrm{M}}} \sqrt{(1-\omega_i)P_i^c}\boldsymbol{f}^c\boldsymbol{H}^{(i,c)}\mathbf{1}_{N_{\mathrm{Tx}}} x_m^- + \sum_{l=N_{\mathrm{M}}+1}^{N} \sqrt{P_l^c}\boldsymbol{f}^c\boldsymbol{H}^{(l,c)}\boldsymbol{W}^{(l,c)}\boldsymbol{\chi}_l + \boldsymbol{f}^c\boldsymbol{n}^c. \tag{8}$$

The post-equalization SINR for RE c can then be calculated as below:

$$\gamma^c = \left(\sum_{i=1}^{N_{\mathrm{M}}} (1-\omega_i)P_i^c \|\boldsymbol{f}^c\boldsymbol{H}^{(i,c)}\mathbf{1}_{N_{\mathrm{Tx}}}\|^2 + \sum_{l=N_M+1}^{N} P_l^c \|\boldsymbol{f}^c\boldsymbol{H}^{(l,c)}\boldsymbol{W}^{(l,c)}\|^2 + \|\boldsymbol{f}^c\|^2\sigma_N^2\right)^{-1} \tag{9}$$

C. AWGN Equivalent SINR

In our analysis, Mutual Information Effective SINR Mapping (MIESM) based PHY abstraction is employed to link system level simulation and link level simulation [17]. Let $f_m(\cdot)$ be the Bit-Interleaved Coded Modulation (BICM) capacity of modulation alphabet m that is associated with the MCS m. Then, across all N_{RE} REs used for the MBSFN TB transmission, the equivalent SINR over AWGN channel for modulation m is:

$$\gamma_m = f_m^{-1}\left[\frac{1}{N_{\mathrm{RE}}}\sum_{c=1}^{N_{\mathrm{RE}}} f_m(\gamma^c)\right]. \tag{10}$$

This AWGN equivalent TB SINR γ_m, together with the AWGN BLER curves, target BLER 1 %, and CQI switching points [10], is used for MCS selection and CQI reporting in system level simulations, as well as packet loss determination when the UE receives a packet.

IV. Simulation Design and Results

A. Simulation Design

The simulation flow chart is illustrated in Figure 3. For comparison purpose, there are two branches, one for unicast and one for MBSFN. Both are run for the same scenario settings. The unicast branch follows Vienna system level simulator [17][1], with TM9 selected. The MBSFN branch utilizes the analytical model derived in Section III. Due to the stringent requirement on transmission reliability in public safety incidents, the minimum MCS among first responders is selected for MBSFN transmissions.

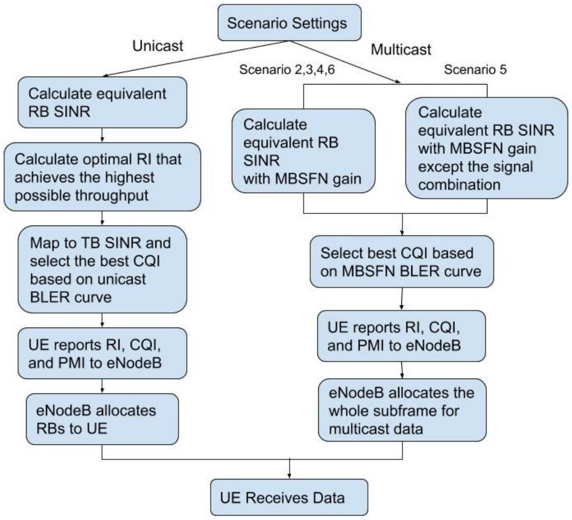

Scenario Settings
Unicast
Multicast
Scenario 2,3,4,6
Scenario 5
Calculate equivalent RB SINR
Calculate optimal RI that achieves the highest possible throughput
Map to TB SINR and select the best CQI based on unicast BLER curve
UE reports RI, CQI, and PMI to eNodeB
eNodeB allocates RBs to UE
Calculate equivalent RB SINR with MBSFN gain
Calculate equivalent RB SINR with MBSFN gain except the signal combination
Select best CQI based on MBSFN BLER curve
UE reports RI, CQI, and PMI to eNodeB
eNodeB allocates the whole subframe for multicast data
UE Receives Data

Fig. 3. Simulation Flowchart

The network simulated is a regular hexagonal network with three sectors per cell, as shown in Figure 2, where the numbers are cell IDs and outer rings are included to simulate interference. Public safety band B14 is selected and consider 8 x 4 MIMO configuration. Channel models used are the urban and rural models defined by 3GPP [18] for path loss, Claussen model for shadowing [19], and International Telecommunication Union (ITU) PedB, VehA, and VehB for small scale fading, with speed 10 km/h, 30 km/h, and 120 km/

[1] Any mention of commercial products in the paper is for information only; it does not imply recommendation or endorsement by National Institute of Standards and Technology.

h, respectively. Inter-site-distance (ISD) studied are 500 m, 1299 m, and 1732 m. Later we use urban ISD1299 VehA to denote urban path loss, ISD 1299 m, and VehA with speed 30 km/h. Other channel notations can be explained similarly. The number of runs and TTIs are chosen to capture statistically stable results.

Table I summarizes the six scenarios simulated for performance analysis, where scenario 5 is designed for MBSFN gain analysis used in Section IV-C and will be described there. The scenarios with different MBSFN area sizes are illustrated in Figure 4, where each color represents a tri-sector site and different color transparency shows different cells within a site. Note that scenarios 2 to 4 have small MBSFN areas to cover public safety incidents with small areas, and scenario 6 has a large enough MBSFN area to cover public safety incidents with large areas.

TABLE I
SIMULATION SCENARIOS

Scenario	Transmission
1	Unicast
2	MBSFN area: cell 1
3	MBSFN area: cell 1-2
4	MBSFN area: cell 1-3
5	MBSFN area: cell 1-21, No Signal Combination
6	MBSFN area: cell 1-21

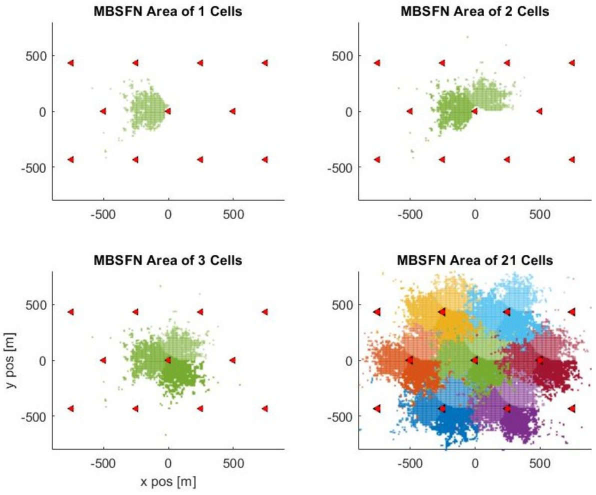


Fig. 4. MBSFN Deployment Scenarios

B. Resource Efficiency and Throughput

In this section we investigate resource efficiency for unicast and MBSFN as well as throughput experienced by first responders. For resource efficiency, the metric studied is information bits/symbol as in Table 7.2.3-1 in [13]. For unicast, in case there is more than one layer from spatial multiplexing, there are two codewords. Information bits/symbol is then the sum of bits/symbol from both codewords.

Scenario 1, 2, 3, 4, and 6 in Table I are simulated for this study, and 10 UEs are dropped uniformly within cell 1. Urban ISD1299 VehA channel is employed, proportional fairness is applied for scheduler, and 50 runs are performed. The resulting cumulative distribution function (CDF) of UE information bits/symbol is shown in Figure 5, where each UE contributes one sample. UE bits/symbol is calculated as the average bits/symbol over total resource units. Specifically, for unicast, it is averaged over all RBs and 300 TTIs. For MBSFN, it is averaged over all TTIs as a single TB is used per TTI (Section II).

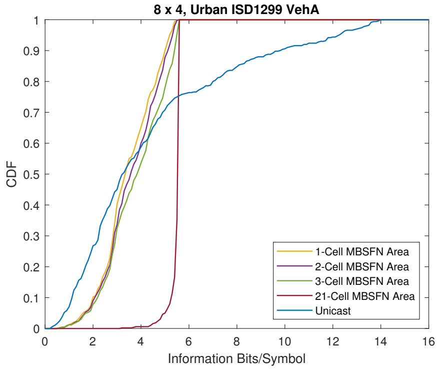


Fig. 5. Unicast vs MBSFN - Information Bits/Symbol

As expected, Figure 5 shows that larger MBSFN size leads to higher bits/symbol. This is because a larger MBSFN area implies more cells participating in MBSFN transmissions, hence the higher diversity combining gain, the less destructive interference, and the higher bits/symbol (Section III).

In addition, Figure 5 also shows that unicast has larger spread. The higher bits/symbol values come from multi-layer-transmissions due to spatial multiplexing in TM9 while the lower values are due to the lack of diversity combining gain.

Nevertheless, the possible higher bits/symbol of unicast does not necessarily imply higher throughput. The CDF of UE throughput from the same simulations is plotted in Figure 6.

Figure 6 shows that although unicast may have higher resource efficiency for individual UEs, MBSFN always has higher throughput. This holds even when the MBSFN area size is one cell. The reason is that in MBSFN each transmission serves all UEs and utilizes all RBs, whereas in unicast each transmission serves one UE and could use only a fraction of all RBs (10 UEs share the total RBs in each run). The more UEs being served, the less RBs each UE gets, and the less resulting throughput. Note that as mentioned in Section II, only 60 % of subframes are used in MBSFN. That is, MBSFN achieves higher throughput with only 60 % of the resources.

Figure 6 also shows that a larger MBSFN area leads to higher throughput. As with the previous analysis, this is because a larger MBSFN area leads to higher diversity combining gain and less destructive interference, and hence higher throughput.

Similar simulations were run for additional channels, and Table II lists the differences in median throughput between MBSFN and unicast, with unicast median throughput as the

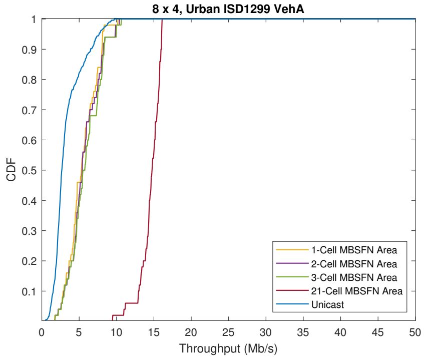


Fig. 6. Unicast vs MBSFN - Throughput

reference point. It can be seen that the trend of higher throughput with larger MBSFN size holds for all channels, and the median throughput delta between unicast and MBSFN can exceed 3 Mb/s.

TABLE II
THROUGHPUT DELTA BETWEEN UNICAST AND MBSFN (MB/S)

MBSFN area size	1 cell	2 cells	3 cells	21 cells
Rural ISD1299 VehB	0.5	0.6	0.6	3.1
Rural ISD1732 VehB	0.6	0.8	1.0	2.8
Urban ISD500 PedB	0.5	0.8	0.9	2.0
Urban ISD500 VehA	0.6	0.8	1.0	2.1
Urban ISD1299 VehA	0.1	0.3	0.6	2.3
Urban ISD1299 VehB	0.1	0.4	0.7	2.3

C. *Sources of MBSFN Gain*

As mentioned in Section II, the performance enhancement of MBSFN mainly comes from two sources: the diversity combining gain from multiple signal sources (cells), and the reduction of inter-cell interference. In this subsection we quantify the contribution of these two sources to MBSFN TB SINR incurred by a random UE (Equation 10). Note that the TB SINR for a UE directly reflects the physical channel experienced by the UE and is used for MCS selection. Hence it reflects the upper bound of the achievable throughput for the UE, and impacts its actual throughput as well.

For this purpose, the test scenario 5 is designed. In its SINR calculations and except for cell 1, power from other MBSFN cells is not included into neither the signal portion nor the interference portion. Hence, compared with scenario 2, inter-cell interference is removed. And compared with scenario 6, there is no diversity combining gain.

All these three scenarios - scenario 2, 5, and 6 - are simulated for 1000 TTIs, and UEs are dropped to cover the entire cell 1 (refer to Figure 4). Figure 7 plots the sorted MBSFN TB SINR in scenario 2, together with MBSFN gains from interference reduction, diversity combining, and the sum of both, respectively. Since the data is sorted by TB SINR, low UE index range maps to cell edge, and high UE index range maps to cell center. It can be seen that from cell edge to cell center, while the diversity combining gain decreases, the gain from interference reduction increases significantly. This is because the power from other MBSFN cells decreases when going from cell edge to cell center. Whereas at cell edge the two gains are comparable, around 5 dB, at cell center the gain from interference reduction dominates and the overall gain could be as high as around 20 dB. Unfortunately, this big SINR improvement at cell center would not lead to significant performance improvement due to the cap of the highest MCS (refer to Table I in [10]) and lack of MIMO technologies (Section II). Note that from the CQI switching points in the table, the maximum CQI 15 can be achieved when SINR is 20.5 dB or above, which can be achieved at cell center without the SINR improvement (the TB SINR in scenario 2). Figure 8 shows MBSFN TB SINR with UE locations, which further verifies the above analysis.

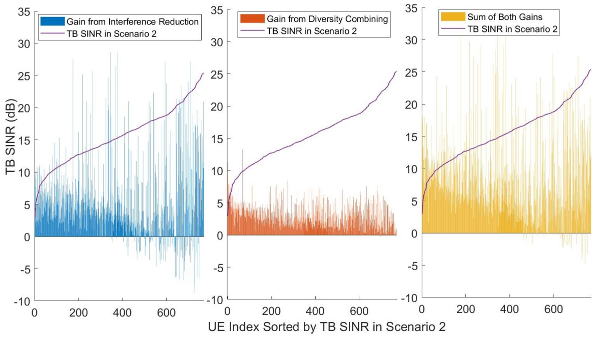


Fig. 7. Sources of MBSFN Gain

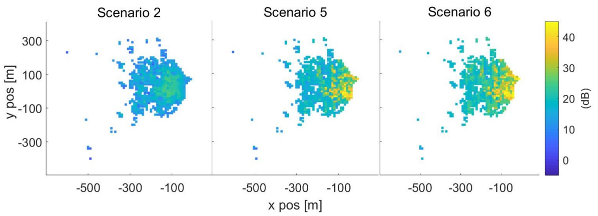


Fig. 8. TB SINR with UE Position

D. *Outage Probability*

In this subsection we quantify improvements in outage probability generated by using MBSFN. Consider a UE dropped randomly at a cell at a random time and being allocated a random resource chunk. In unicast, the resource chunk considered is five RBs, while in MBSFN it is one subframe. The outage probability considered in this paper is defined as the probability that the information bits/symbol of this UE is below a threshold.

To capture this outage probability, for each scenario listed in Table I except scenario 5, UEs are dropped to cover the entire cell 1, and simulations are run for 2000 TTIs. The resulting outage probability as a function of the threshold is plotted in Figure 9. The figure shows a significant decrease in outage probability when using MBSFN for all MBSFN sizes studied. As with the previous analysis, this is because of diversity combining and interference reduction in MBSFN. The larger the MBSFN area size, the smaller the outage probability is. For example, when the threshold is 1 bit/symbol, the outage probability for unicast is around 4.5 %, while for MBSFN it is less than 0.2 % for MBSFN area sizes of 1 to 3 cells, and 0 % for 21 cells. Note that the significant decrease in outage probability holds even for an MBSFN area size of one cell.

Table III shows the achievable information bits/symbol when the outage probability is 2 % and 5 %, respectively. With the same planned outage probability, the table also shows a significant increase in achievable bits/symbol from unicast to MBSFN. Due to quantization, the bits/symbol is the same with MBSFN area size of 1, 2, and 3 cells. While with area size of 21 cells, the achievable bits/symbol is significantly higher.

TABLE III
INFORMATION BIT/SYMBOL WITH OUTAGE PROBABILITY

Scenario	1	2	3	4	6
2 %	0.83	1.91	1.91	1.91	3.32
5 %	1.17	1.47	1.47	1.47	4.52

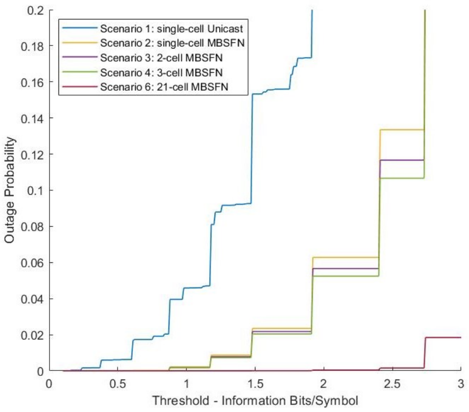


Fig. 9. Outage Probability with Threshold

V. CONCLUSION

To meet first responders' communication requirements, in this paper LTE MBSFN and unicast were investigated from the perspectives of throughput, information bits/symbol, sources of MBSFN gain, and outage probability. First, we derived an accurate MBSFN SINR analytical model. Then, through comprehensive MBSFN and unicast simulations, we showed that although the physical channels in unicast may be more efficient with higher information bits/symbol, MBSFN would provide higher throughput since the resources are not shared among UEs as in unicast. We also validated that a larger MBSFN area size would increase the MBSFN throughput due to higher MBSFN diversity combining gain and less interference. We further demonstrated that SINR improvement due to MBSFN diversity combining gains and interference reduction are comparable at cell edge, while the high MBSFN SINR improvement at cell center does not lead to significant performance improvement due to MCS cap and lack of MIMO support for MBSFN. Finally, compared with unicast, MBSFN improves outage probability significantly, which is essential for public safety incidents. All the above results hold not only for large MBSFN area sizes as in commercial networks, but also for small MBSFN area sizes, which cover cases where the incident area is small.

Since resource sharing is one major factor that limits unicast performance, our next step is to investigate the impact of number of first responders on MBSFN performance improvement over unicast. In addition to throughput, we will introduce other performance metrics as well, such as flight time, and further investigate the trade-offs between unicast and MBSFN.

REFERENCES

[1] "Minnesota Department of Public Safety Public Safety Wireless Data Network Requirements Project Needs Assessment Report," Minnesota PSN, Tech. Rep., May, 2011.

[2] J. Song and R. Phung, "Emergency group call over embms," in *16th International Conference on Advanced Communication Technology*, Feb 2014, pp. 1017–1022.

[3] T. Doumi, M. F. Dolan, S. Tatesh, A. Casati, G. Tsirtsis, K. Anchan, and D. Flore, "Lte for public safety networks," *IEEE Communications Magazine*, vol. 51, no. 2, pp. 106–112, February 2013.

[4] R. Rebhan and J. Zander, "On the outage probability in single frequency networks for digital broadcasting," *IEEE Transactions on Broadcasting*, vol. 39, no. 4, pp. 395–401, Dec 1993.

[5] G. Malmgren, "On the performance of single frequency networks in correlated shadow fading," *IEEE Transactions on Broadcasting*, vol. 43, no. 2, pp. 155–165, June 1997.

[6] L. Rong, O. B. Haddada, and S. Elayoubi, "Analytical analysis of the coverage of a mbsfn ofdma network," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, Nov 2008, pp. 1–5.

[7] F. X. A. Wibowo, A. A. P. Bangun, A. Kurniawan, and Hendrawan, "Multimedia Broadcast Multicast Service over Single Frequency Network (MBSFN) in LTE based Femtocell," in *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics*, July 2011, pp. 1–5.

[8] S. Mitrofanov, A. Anisimov, and A. Turlikov, "eMBMS LTE Usage to Deliver Mobile Data," in *2014 6th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, Oct 2014, pp. 60–65.

[9] J. Chen, M. Chiang, J. Erman, G. Li, K. K. Ramakrishnan, and R. K. Sinha, "Fair and Optimal Resource Allocation for LTE Multicast (eMBMS): Group Partitioning and Dynamics," in *2015 IEEE INFOCOM*, April 2015, pp. 1266–1274.

[10] C. Liu, C. Shen, J. Chuang, A. R. Rouil, and H. Choi, "Throughput Analysis between Unicast and MBSFN from Link Level to System Level," in *IEEE 90th Vehicular Technology Conference*, September 2019.

[11] FirstNet. https://firstnet.gov/content/firstnet-building-nationwide-public-safety-network.

[12] J. Zhang, L. Yang, L. Hanzo, and H. Gharavi, "Advances in Cooperative Single-Carrier FDMA Communications: Beyond LTE-Advanced," *IEEE Communications Surveys Tutorials*, vol. 17, no. 2, pp. 730–756, Second quarter 2015.

[13] 3GPP TS36.213, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures," 3GPP, Standard, Jan. 2019.

[14] 3GPP TS36.211, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," 3GPP, Standard, Dec. 2018.

[15] S. Sesia, I. Toufik, and M. Baker, *LTE - The UMTS Long Term Evolution: From Theory to Practice: Second Edition*, August 2011.

[16] 3GPP TS36.300, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2," 3GPP, Standard.

[17] M. Rupp, S. Schwarz, and M. Taranetz, *The Vienna LTE-Advanced Simulators: Up and Downlink, Link and System Level Simulation*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[18] 3GPP TS36.942, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Frequency (RF) system scenarios," 3GPP, , Jul. 2018.

[19] Claussen, "Efficient modelling of channel maps with correlated shadow fading in mobile radio systems," in *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, Sep. 2005, pp. 512–516.

DesignCon 2020

DARPA Organic Interposer Characterization

Dylan Williams, NIST
dylan.williams@nist.gov

Richard Chamberlin, NIST
richard.chamberlin@nist.gov

Jerome Cheron, NIST
jerome.cheron@nist.gov

Brent DeVetter
brent.devetter@gmail.com

Sam Chitwood
samchitwood@gmail.com

Ken Willis, Cadence
kenw@cadence.com

Brad Butler, Cadence
brad@cadence.com

Farhang Yazdani, Broadpak Corp.
farhang.yazdani@broadpak.com

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

Abstract

We report on a study of interconnects fabricated on organic and silicon interposers used to connect state-of-the art digital, analog and RF chiplets commissioned by the U.S. Defense Advanced Research Projects Agency (DARPA). The interconnects were characterized with state-of-the-art on-wafer measurement methods developed at the National Institute of Standards and Technology (NIST) and then simulated with the Cadence® Sigrity^{TM} simulation software package. The two-port, four-port and eight-port measurements and calibrations were performed to frequencies as high as 110 GHz using custom on-wafer calibrations to improve accuracy. We discuss the measurement and simulation methodologies and present detailed comparisons of the measurement and simulation results.

Commercial brand names are used only to better explain the experiments and analysis. The National Institute of Standards and Technology does not endorse commercial products. Other products may work as well or better.

This work was supported by the DARPA Microsystems Technology Office (MTO) under the DARPA Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) Program. The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Reference herein to any specific commercial product, process, or service by trade name, trademark or other trade name, manufacturer or otherwise, does not necessarily constitute or imply endorsement by DARPA, the Defense Department or the U.S. government, and shall not be used for advertising or product endorsement purposes.

Approved for Public Release, Distribution Unlimited

Biographies

Dylan F. Williams received a Ph.D. in Electrical Engineering from the University of California, Berkeley in 1986. He joined the Electromagnetic Fields Division of the National Institute of Standards and Technology in 1989 where he develops electrical waveform and microwave metrology. He has published over 80 technical papers and is a Fellow of the IEEE. He is the recipient of the Department of Commerce Bronze and Silver Medals, the Astin Measurement Science Award, two Electrical Engineering Laboratory's Outstanding Paper Awards, three Automatic RF Techniques Group (ARFTG) Best Paper Awards, the ARFTG Automated Measurements Technology Award, the IEEE Morris E. Leeds Award, the 2011 European Microwave Prize and the 2013 IEEE Joseph F. Keithley Award. Dylan also served as Editor of the IEEE Transactions on Microwave Theory and Techniques from 2006 to 2010 and as the Executive Editor of the IEEE Transactions on Terahertz Science and Technology.

Richard A. Chamberlin graduated from the University of California (Santa Barbara, CA) with a B.S. in physics in 1984, and obtained his Ph.D. in physics from the Massachusetts Institute of Technology (Cambridge, MA) in 1991. In 1995 he was the first winter-over scientist with the pioneering Antarctic Submillimeter Telescope and Remote Observatory which he helped design, build, and test while at Boston University. From 1996 to 2010 he was the Technical Manager of the Caltech Submillimeter Observatory located on the summit of Mauna Kea on the Island of Hawaii. Currently he is working in the High Speed Electronics Group at the National Institute of Standards and Technology in Boulder, Colorado.

Jerome Cheron received the Ph.D. degree in electrical engineering from the University of Limoges, France, in 2011. His research project was led at XLIM laboratory and Thales Air Systems, France. In 2012, he was Postdoctoral Fellow with the Fraunhofer IAF, Freiburg, Germany. In 2013, he joined NIST in Boulder, Colorado. His research interests include the design and optimization of microwave and millimeter-wave circuits in III-V technology.

Brent DeVetter received his Ph.D. in Electrical Engineering with a focus on plasmonic materials from the University of Illinois at Urbana-Champaign in 2016. He was a postdoctoral researcher at NIST working in the High-Speed Electronics Group performing high frequency measurements and electromagnetic simulations. He also completed a postdoctoral appointment at the Pacific Northwest National Laboratory focusing on infrared and terahertz spectroscopy of crystalline materials.

Sam Chitwood has focused on signal and power integrity analysis for 17 years. He has been a product engineer at Cadence Design Systems and an application engineer with Sigrity. He contributed to the DARPA CHIPS project through October 2019 while at Cadence. He has BSEE and MSECE degrees from Georgia Institute of Technology.

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

Ken Willis is a Product Engineering Group Director focusing on system-level analysis solutions at Cadence Design Systems. He has over 30 years of experience in the modeling, analysis, design, and fabrication of high-speed digital circuits. Prior to Cadence, Ken held engineering, technical marketing, and management positions with the Tyco Printed Circuit Group, Compaq Computers, Sirocco Systems, Sycamore Networks, and Sigrity.

Brad Butler is a Senior Application Engineer at Cadence Design Systems, focused on enabling advanced electronic systems design in the Aerospace and Defense sector. Prior to joining Cadence, he worked on vehicle dynamics, autonomous driving, and LIDAR applications at a small automotive startup. Brad earned B.S. degrees in Mechanical Engineering and Economics from North Carolina State University.

Farhang Yazdani is the President and CEO of BroadPak Corporation. Through his 20 years with the industry, he has served in various technical, management, and advisory positions with leading semiconductor companies worldwide. He is the author of the book "Foundations of Heterogeneous Integration: An Industry-Based, 2.5D/3D Pathfinding and Co-Design Approach". He is the recipient of 2013 NIPSIA award in recognition of his contribution to the advancement and innovations in packaging technologies. He has numerous publications and IPs in the area of 2.5D/3D Packaging and Assembly, serves on various technical committees and is a frequent reviewer for IEEE Journal of Advanced Packaging. He received his undergraduate and graduate degrees in Chemical Engineering and Mechanical Engineering from the University of Washington, Seattle.

I. Introduction

Large, monolithic ICs are being replaced by arrays of chiplets mounted on advanced interposers (passive integrated circuits supporting complex interconnects between the die mounted on them) in wide ranging applications of our industry. A few high-profile examples are Intel's "Foveros" and EMIB interposers [1, 2] and AMD's latest Ryzen processors [3, 4]. Chiplet-to-chiplet interconnects with micron and sub-micron widths pose new design, analysis and implementation challenges for digital and analog designers.

DARPA has supported work in heterogenous integration at RF frequencies though the Diverse Accessible Heterogeneous Integration (DAHI) Program [5-9] and this has helped lead a larger push for heterogenous integration at higher frequencies [10]. The National Institute of Standards and Technologies (NIST) performed work under this program characterizing heterogenous interconnects between several different technologies [11].

More recently DARPA has expanded this thrust to digital and analog integrated circuits with the Common Heterogeneous Integration and IP Reuse Strategies (CHIPS) Program [12, 13] under which this work was performed. In this paper, we report on a study of interconnects fabricated on build-up and silicon interposers used to connect state-of-the art digital, analog and RF chiplets supported by this program.

The study performed under the DARPA CHIPS Program focused on two interconnect types, an organic interposer supporting an Ajinomoto GX-92 build-up film (ABF) interconnect [14] covered with Hitachi SR7300 solder resist [15] and an inorganic silicon interposer with through-substrate vias (TSVs) supporting a silicon-dioxide (SiO_2) thin-film interconnect. The organic interposer provides an inexpensive but high-performance solution for integrating chiplets while the inorganic silicon interposer represents the state-of-the-art in interposer technologies. The interposers, and the calibration kits and test structures required to accurately characterize them, were designed and fabricated under the program for the purpose of obtaining the most accurate technology characterization possible to frequencies up to 110 GHz.

The goal of this paper is to study the accuracy of the Cadence® Sigrity^{TM} simulation software package, which includes an efficient hybrid solver for efficient S-parameter extraction of large interconnect structures and a slower but more accurate full-wave solver for more complicated 3-D structures. We first performed the simulations of the test structures on the interposer with the Sigrity hybrid solver using dimensional and other measurements performed by the manufacturer. Then, we compared the measurement and hybrid-solver simulation results and assessed the ability of the hybrid solver to predict interconnect performance in the two technologies we studied. Finally, in cases where we observed significant disagreement, we re-ran the simulations using the Sigrity full-wave solver. This allowed us to better assess when the faster hybrid solver was sufficient and when the slower but more accurate full-wave solver was needed.

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

II. Measurements

Most on-wafer measurements are calibrated with an “impedance-standard substrate” fabricated by the probe manufacturer. These are usually referred to as “probe-tip” calibrations. Probe-tip calibrations are performed under the assumption that the interaction between the probe tips and the contact pads on the wafer can be safely ignored. This is generally the case at lower microwave frequencies. However, the electrical parasitics associated with the interaction between the probe tips, contact pads, and even the device under test, increase roughly linearly with frequency [16-18]. For example, the admittance associated with the capacitance of the contact pads used to connect to a device will typically increase linearly with the frequency, causing measurement errors to increase with frequency as well [16-18]. There are more subtle errors introduced into probe-tip calibrations as well. For example, the inductance associated with a short bar printed on an impedance-standard substrate increases with the probe pitch, making it difficult to define the impedance of the short fabricated on the impedance-standard substrate. As the errors introduced into the calibrations are smooth and roughly linear, these errors are often not immediately apparent, but can nevertheless grow quite large at higher frequencies. For these reasons, the most accurate microwave calibrations are performed in transmission lines.

II.1. The Thru-Reflect-Line Calibration

The goal of a probe-tip calibration is to place the calibration and measurement reference planes at the tips of the probes, again with the assumption that the electrical parasitics associated with the interaction of the probe tips, contact pads and device under test can be ignored. The goal of an on-wafer thru-reflect-line (TRL) calibration, on the other hand, is to place the calibration and measurement reference planes in a transmission line fabricated on an integrated circuit or interconnect. In this environment, voltages, currents, scattering parameters and other electrical quantities can be rigorously defined (and measured) even at microwave frequencies [19-23], something that cannot be said of probe-tip calibrations. Furthermore, the result of cascading devices at these reference planes can be predicted with rigor from measurements performed at these reference planes [19]. Finally, the calibration and measurement reference plane can be moved to very close to the device under test, eliminating many parasitics in the measurement of the device. This generally proves more effective than conventional parasitic-extraction techniques used in transistor modeling, for example, as parasitics are first minimized before being estimated [24, 25].

An on-wafer microstrip TRL calibration kit fabricated in the interconnect stack of a silicon die is illustrated in Fig. 1 below. The TRL calibration kit includes a short “thru” line between the port 1 and port 2 contact pads, a number of longer lines, and a symmetric reflect (usually a short section of transmission line on each port terminated with an electrical open or an electrical short). These calibration standards are enough to perform a calibration in the transmission lines with a reference plane at the center of the thru line [19, 23].

Figure 2 illustrates the standards, their definitions (*i.e.* their properties), and their functions in the TRL calibration. The thru sets most of the calibration coefficients. The symmetric reflect allows the calibration to set the reference plane in the center of the thru. The lines set the reference impedance of the calibration to the characteristic impedance Z_0 of the transmission line and provide a measurement of the propagation constant γ of the transmission line.

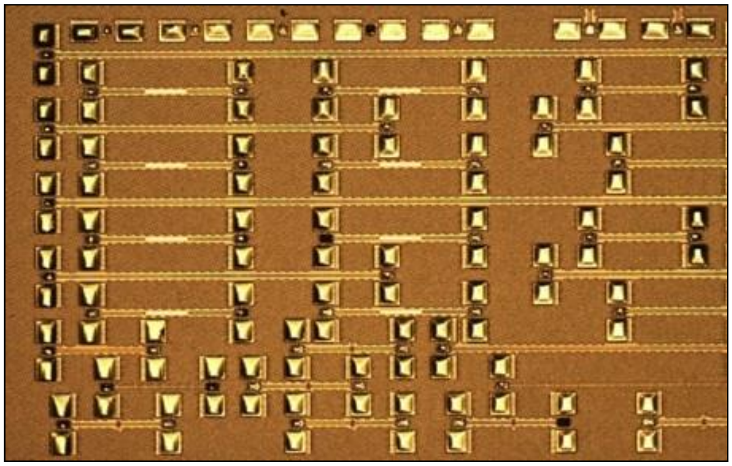
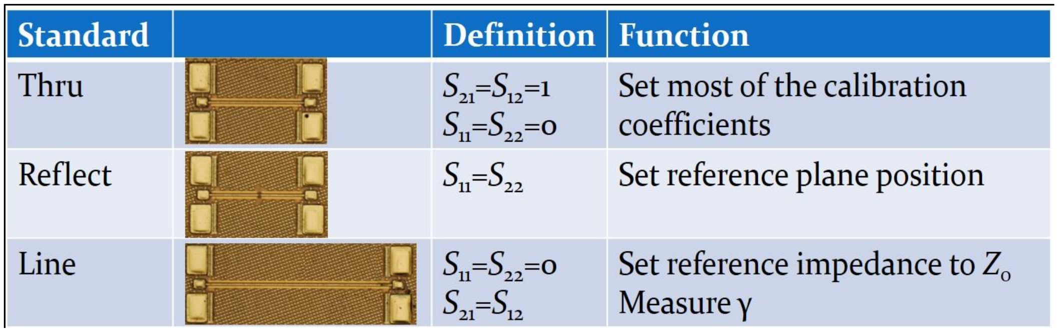


Figure 1. TRL calibration kit fabricated on a silicon die.



Standard		Definition	Function
Thru		$S_{21}=S_{12}=1$ $S_{11}=S_{22}=0$	Set most of the calibration coefficients
Reflect		$S_{11}=S_{22}$	Set reference plane position
Line		$S_{11}=S_{22}=0$ $S_{21}=S_{12}$	Set reference impedance to Z_0 Measure γ

Table 1. TRL calibration standards.

II.2. SiO₂ Transmission-Line Characteristic Impedance

Calibrations with a reference impedance equal to the characteristic impedance of a transmission line are not usually very useful, as the characteristic impedance is generally complex and frequency dependent [23]. Thus, the next step of the TRL calibration is to determine the characteristic impedance of the calibration.

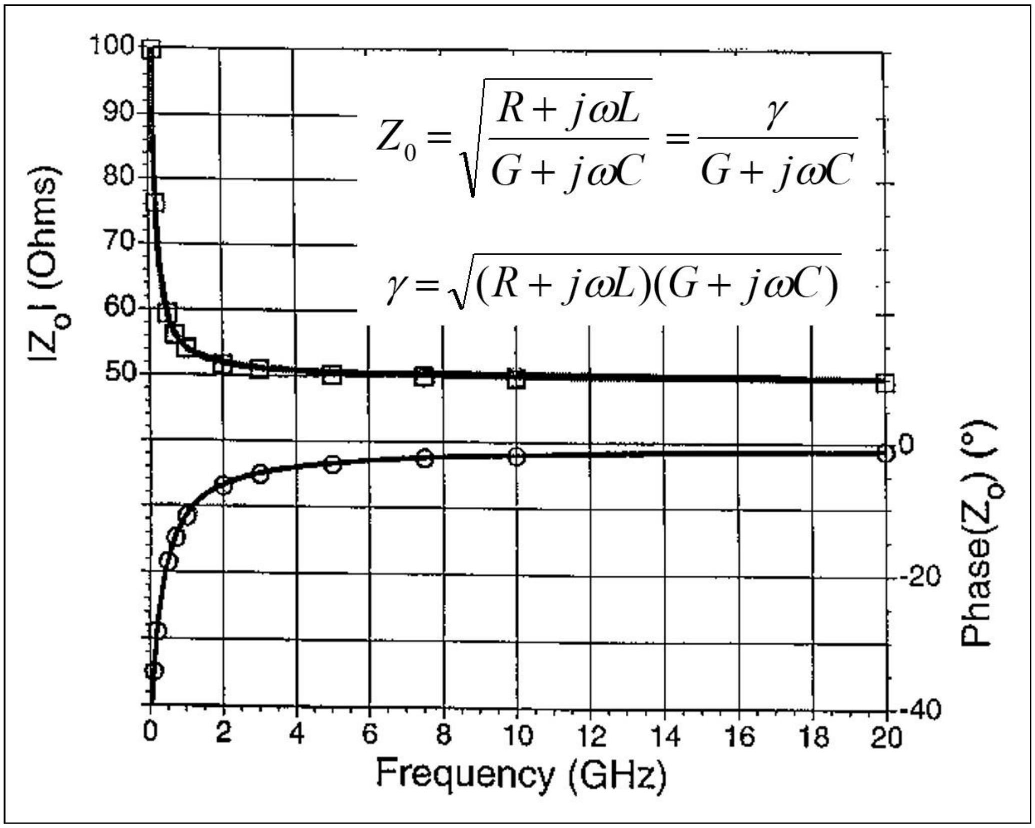


Figure 2. Measurement of the characteristic impedance Z_0 of a coplanar waveguide transmission line compared to calculation. The measurements are shown in solid lines. The calculated magnitude of Z_0 is shown in squares and the calculated phase of Z_0 is shown in circles. The formulas in the upper right were used to calculate Z_0. From [27].

We had two interconnect types to consider, the organic interconnect fabricated on a woven fiberglass core and the inorganic SiO_2 interconnect fabricated on a bulk silicon interposer. The SiO_2 has a roughly frequency-independent dielectric constant and a small loss tangent while the electrical behavior of the organic ABF is complicated by low-frequency dielectric relaxation. With this in mind, we start with the more straightforward analysis of the characteristic impedance of the microstip and stripline transmission lines fabricated in the SiO_2 interconnects.

Finding the characteristic impedance is quite straight forward in quasi-TEM transmission lines constructed from low-loss dielectrics, as the per-unit-length capacitance of the transmission line is very nearly frequency independent and the per-unit-length conductance of the transmission line is small. In this case, the capacitance of the transmission line can be determined from the measurement of a small resistor or the measurement of the per-unit-length DC resistance of the line and knowledge of the propagation constant γ estimated by the TRL calibration algorithm [26]. For these quasi-TEM lines fabricated on well-behaved dielectrics, the per-unit-length capacitance of the lines can even be easily determined by field solvers [26].

Once the per-unit-length capacitance C of the quasi-TEM transmission line has been determined, the characteristic impedance of the transmission line can be determined from the formulas in the upper-right of Fig. 2 [27]. In essence, the per-unit-length capacitance *C* and conductance *G* and the propagation constant γ of the transmission line are enough to determine the per-unit-length resistance *R* and inductance *L* of the transmission line. Once *R* and *L* have been determined, everything is known about the line, and the characteristic impedance Z_0 of the transmission line can be calculated directly with the upper formula in Fig. 2.

Figure 2 compares measurements (solid lines) of the magnitude and phase of the characteristic impedance Z_0 of a coplanar waveguide line to simulations (squares and circles) calculated with the full-wave method of [28], which includes field penetration into the metal in a rigorous way. As can be seen from the figure, the agreement is excellent. We used this approach to determine the dielectric constant of the microstrip and stripline transmission lines we used in our TRL calibrations on the SiO_2 interconnects fabricated on the bulk silicon. We determined the per-unit-length capacitance of the transmission lines from measurements of a small resistors fabricated in the interconnects, as described in [26].

II.3. ABF Transmission-Line Characteristic Impedance

The approach to determining the capacitance and conductance of transmission lines fabricated on low-loss inorganic dielectrics described in the last section cannot account for low-frequency transmission-line behavior due to dielectric relaxation. Therefore, the approach described in the last section is not applicable to the microstrip and stripline transmission lines we fabricated on the organic ABF interconnects. Instead, we first estimated the frequency-dependent dielectric constant and loss tangent of the ABF and then estimated the capacitance and conductance of the ABF transmission lines from their dimensions. Finally, we used capacitive parallel-plate test structures to check our result.

We estimated the frequency-dependent dielectric constant and loss tangent of the ABF from measurements of the ABF film performed at 1 GHz by the manufacturer and the low frequency and high frequency limits of the dielectric constant. We then applied the Djordjevic-Sarkar algorithm to extract a dielectric-relaxation model [29]. This model was independently suggested by Svensson in [30]. Figures 3 and 4 on the next page compare the extracted model to the measurements provided by the manufacturer.

We then estimated the capacitance and conductance of the ABF transmission lines from the model for the dielectric constant we determined and from the dimensions of the transmission lines provided by the manufacturer. This was enough to estimate the characteristic impedance Z_0 of the ABF transmission lines using the formulas in Fig. 2.

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

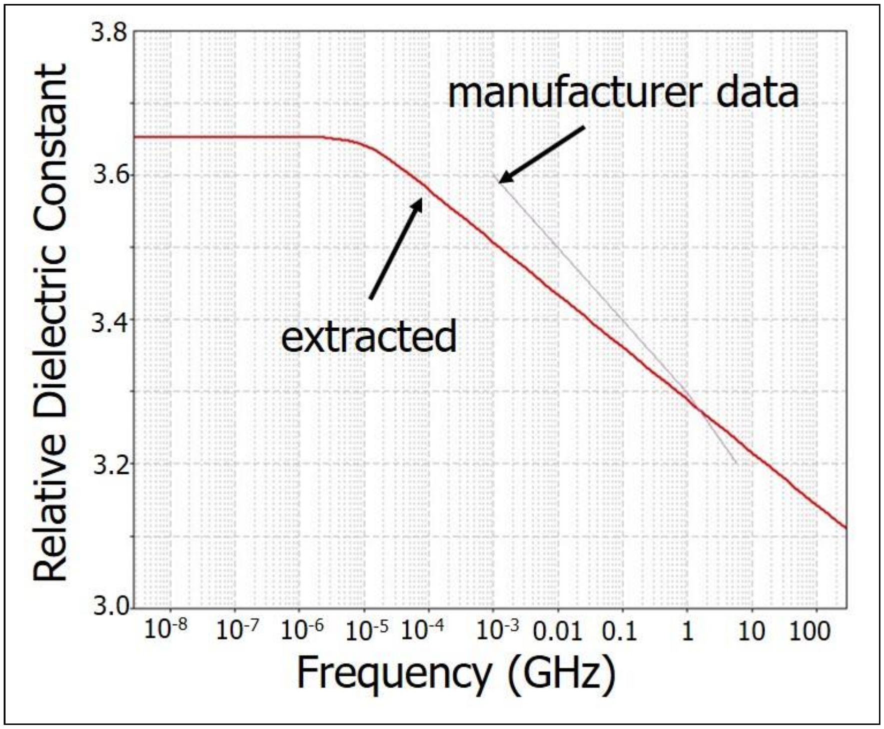


Figure 3. Comparison of the extracted relative dielectric constant and the measurements provided by the ABF manufacturer.

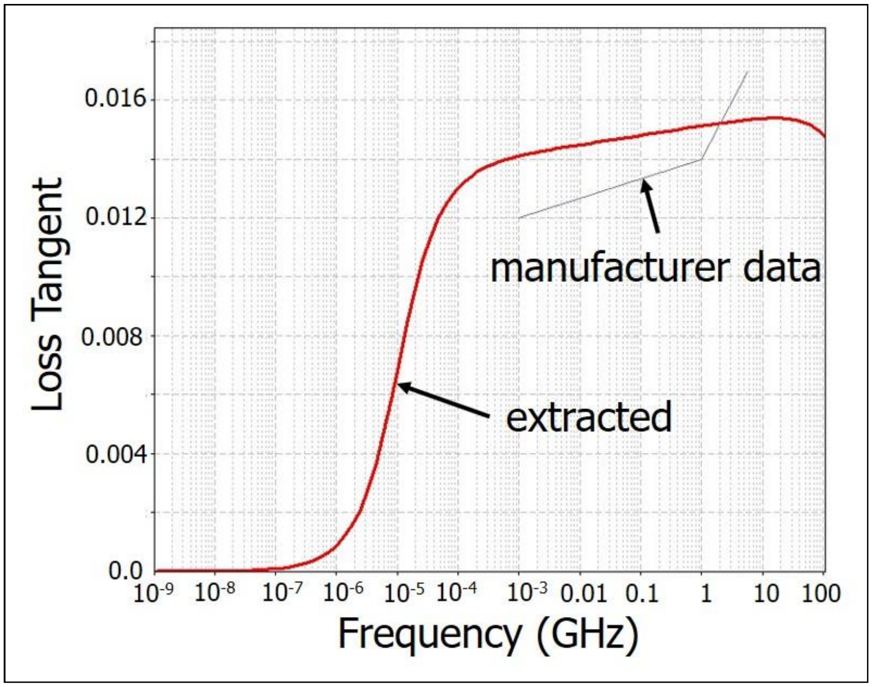


Figure 4. Comparison of the extracted loss tangent and the measurements provided by the ABF manufacturer.

Finally, we verified the dielectric constants and loss tangent we extracted for the organic ABF with large parallel-plate capacitors similar to those routinely created in the power distribution network of the interconnect. The positive plate of each capacitor was sandwiched between an upper and lower ground plate in the interconnect stackup. The two ground plates were stitched together periodically with vias that passed through small holes in the positive plate of the capacitor. We also used via stacks to connect the capacitor's positive plate and the two ground plates to contact pads on the top surface of the interconnect stackup with vias.

To test the capacitors, we first performed a TRL calibration on a fused silica wafer and moved the reference plane of the calibration back to the probe tips. A comparison of the measured and simulated magnitudes of the impedance of the capacitor at the measurement reference plane on the top surface of the ABF interconnect is shown in Fig. 5. While the calibration and extraction approach were both somewhat approximate, they clearly proved adequate for this purpose.

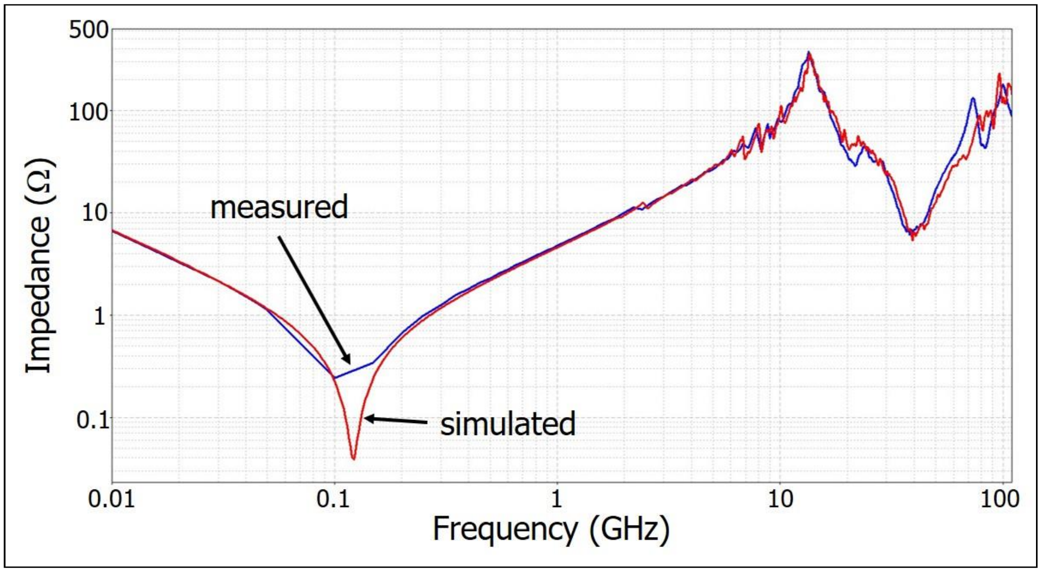


Figure 5. The measured and simulated impedances of one of the parallel-plate capacitors we used to verify extraction procedure we used, as implemented in the Cadence Sigrity software package.

III. Correlation Between Measurements and Simulations

We performed hundreds of TRL measurements and Sigrity hybrid-solver and full-wave solver simulations of test structures fabricated on the organic and silicon interposers and carefully examined the correlation between measurement and simulation. In most cases we were able to perform five or more independent calibrations and measurements of each test structure on different die, which allowed us to gauge the degree of correlation we obtained with the TRL measurements and Sigrity simulations. Here we present a few of these results on the organic ABF interposer performed with the Sigrity hybrid solver to illustrate the high degree of correlation we found between the two.

III.1. Coupled Lines on Organic ABF Interposer

Figure 6 below shows the layout of the organic ABF interposer. The design included microstrip and stripline TRL calibration kits shown in the upper and right-most sections of the layout, as well as a number of 2-port, four-port and eight-port test structures on other parts of the designs. Only the microstrip test structures fabricated in the top-metal levels are visible, but many more stripline test structures are buried in the interconnect stack and were also available for test.

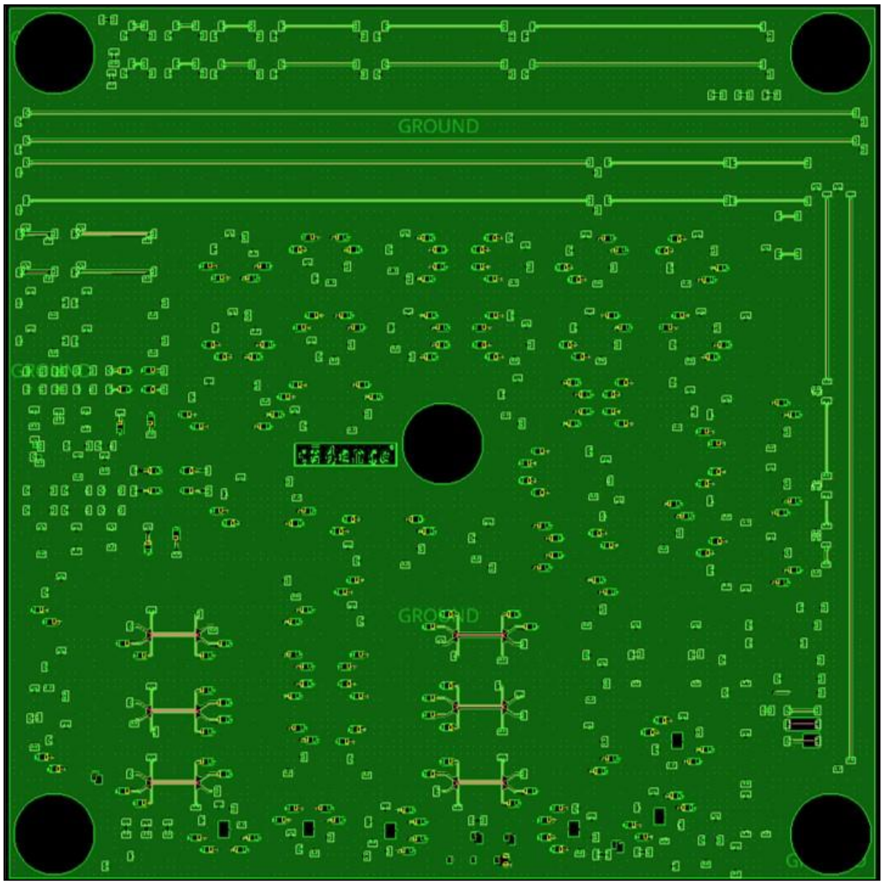
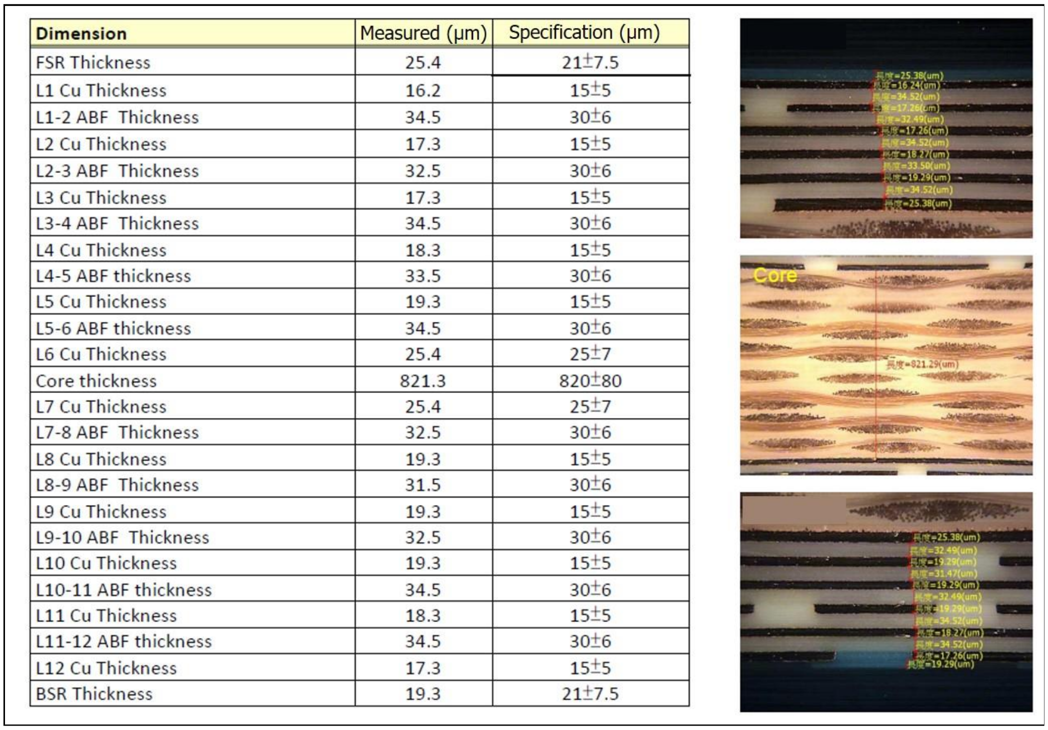


Figure 6. Top layer of the organic ABF Interposer layout.

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

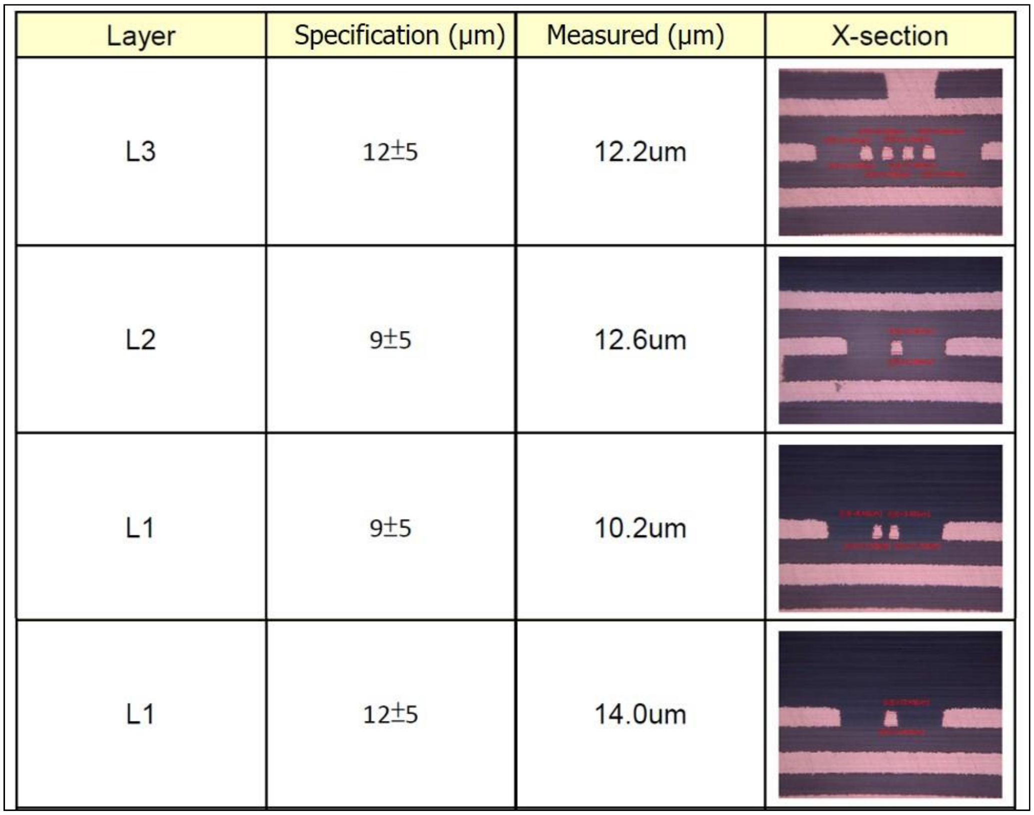
Table 2 shows the twelve-metal ABF interposer stackup. The first column in the table lists the layer names and materials, the second column lists the layer thicknesses measured by the manufacturer and the last column lists the manufacturer's specifications. The photos in the right of the figure are of actual cross sections of the interposers we designed and tested.



Dimension	Measured (μm)	Specification (μm)
FSR Thickness	25.4	21±7.5
L1 Cu Thickness	16.2	15±5
L1-2 ABF Thickness	34.5	30±6
L2 Cu Thickness	17.3	15±5
L2-3 ABF Thickness	32.5	30±6
L3 Cu Thickness	17.3	15±5
L3-4 ABF Thickness	34.5	30±6
L4 Cu Thickness	18.3	15±5
L4-5 ABF thickness	33.5	30±6
L5 Cu Thickness	19.3	15±5
L5-6 ABF thickness	34.5	30±6
L6 Cu Thickness	25.4	25±7
Core thickness	821.3	820±80
L7 Cu Thickness	25.4	25±7
L7-8 ABF Thickness	32.5	30±6
L8 Cu Thickness	19.3	15±5
L8-9 ABF Thickness	31.5	30±6
L9 Cu Thickness	19.3	15±5
L9-10 ABF Thickness	32.5	30±6
L10 Cu Thickness	19.3	15±5
L10-11 ABF thickness	34.5	30±6
L11 Cu Thickness	18.3	15±5
L11-12 ABF thickness	34.5	30±6
L12 Cu Thickness	17.3	15±5
BSR Thickness	19.3	21±7.5

Table 2. ABF interposer stackup with measured layer thicknesses, specifications and cross sections. FSR = front-side solder resist, L*n* = metal layer *n*, L*n*-*m* = ABF layer thickness between L*n* and L*m*, BSR = back-side solder.

Table 2 shows some additional cross sections and width measurements performed by the manufacturer. As can be seen in the tables in Tables 2 and 3, the tolerances on the dimensional specifications were quite large and many of the measured transverse dimensions differed significantly from the nominal specifications. Because of the large variations in manufactured line width and no way to measure physical line widths for specific signals under test, this parameter was used as a "tuning parameter" in simulations. That is, with all other material and geometry parameters held constant, the line width was varied across the tolerance to observe the best match against the measured results. We optimized all of the simulated results we will present below in this way.



Layer	Specification (µm)	Measured (µm)	X-section
L3	12±5	12.2um	
L2	9±5	12.6um	
L1	9±5	10.2um	
L1	12±5	14.0um	

Table 3. Trace widths for selected structures fabricated on the ABF interposers. Spec = designed width for selected structure. X-section data = measured width for selected structure.

Correlations between the 2-port, 4-port and 8-port test structures we designed and tested, and simulations performed with the Sigrity hybrid solver (the faster of the two Sigrity solvers), were generally excellent. We present data here only for a single four-port and a single eight-port device.

III.2. Four-Port Coupled Lines on Organic ABF Interposer

Figure 7 shows the 4-port layout we used to assess measurement and simulations of coupled microstrip and striplines. We used the NIST Microwave Uncertainty Framework [31] to perform the four-port calibrations. This software package allows a lateral “east-west” two-port TRL calibration and a vertical “north-south” two-port TRL calibration to be combined with the use of an additional unknown bend to form a rigorous four-port calibration. We also took advantage of the Uncertainty Framework to move the calibration and measurement reference planes through the microstrip lines and vias, placing them in the striplines where the coupled structures buried in the ABF interconnect stack began and ended. We did not take advantage of the ability of this software package to evaluate measurement uncertainty in this work.

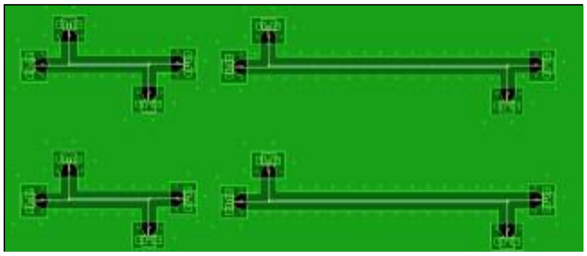


Figure 7. 4-port coupled-line layout.

The following figures illustrate the high-degree of correlation we observed between measurements and simulations performed on two coupled 9 µm wide 1 mm long coupled striplines spaced 9 µm apart apart and buried in the third level of metal in the ABF interconnect stack. Simulations are shown in thick solid lines and measurements performed on other die are shown in thinner lines.

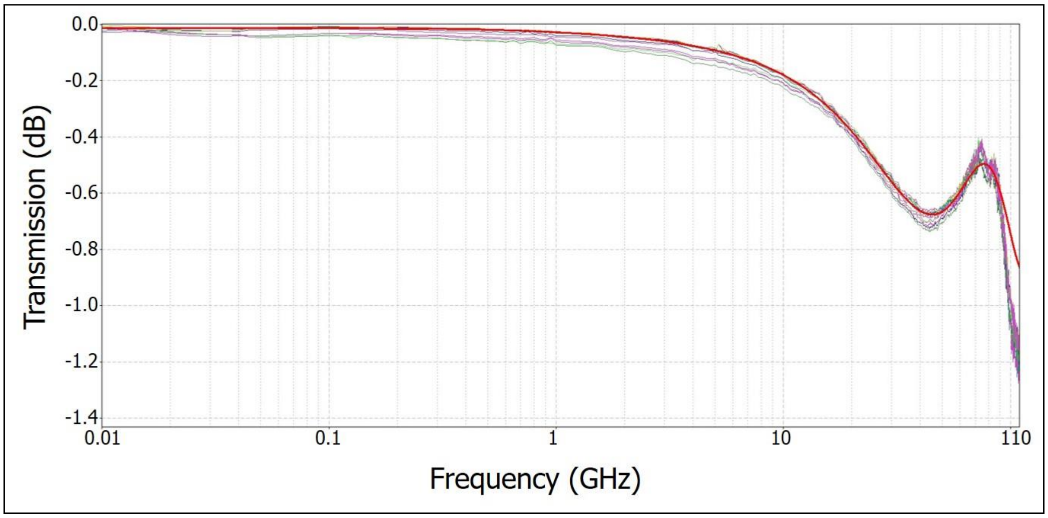


Figure 8. Transmission coefficient through a single 9 µm wide and 1 mm long stripline coupled to an adjacent line 9 µm apart and buried in the third level of metal in the ABF interconnect stack. Simulations are shown in thick solid lines and measurements performed on other die are shown in thinner lines. The hybrid solver was used for the simulations and only the line width was adjusted in the simulation.

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

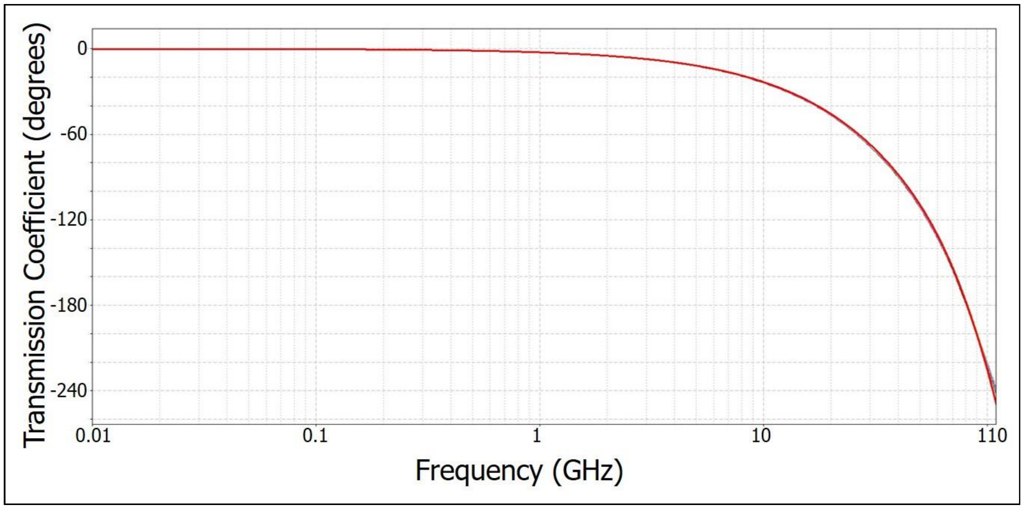


Figure 9. Phase of the transmission coefficient through a single 9 µm wide and 1 mm long stripline coupled to an adjacent line 9 µm apart and buried in the third level of metal in the ABF interconnect stack. Simulations are shown in thick solid lines and measurements performed on other die are shown in thinner lines. The hybrid solver was used for the simulations and only the line width was adjusted in the simulation. (Deviations of measurements are too small to be seen.)

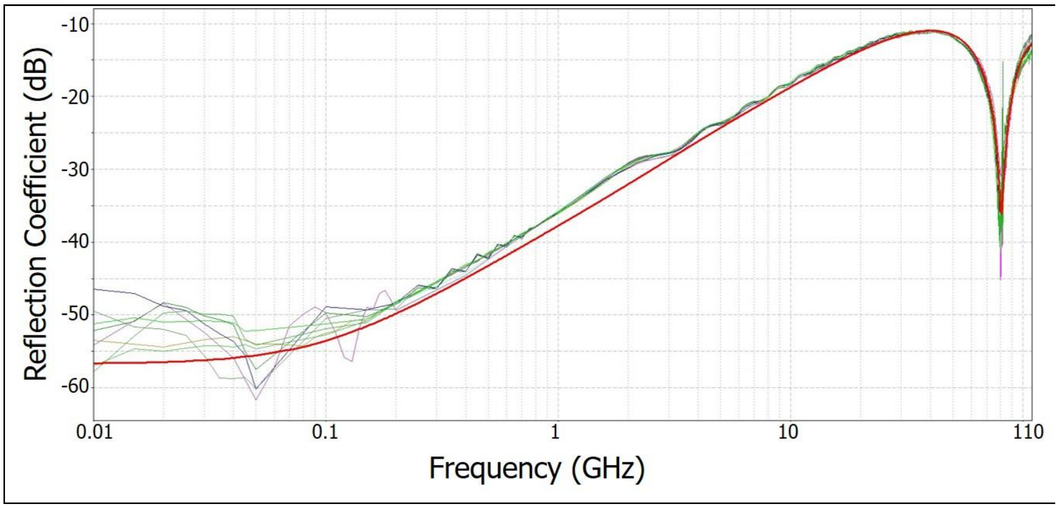


Figure 10. Reflection coefficient of a single 9 µm wide and 1 mm long stripline coupled to an adjacent line 9 µm apart and buried in the third level of metal in the ABF interconnect stack. Simulations are shown in thick solid lines and measurements performed on other die are shown in thinner lines. The hybrid solver was used for the simulations and only the line width was adjusted in the simulation.

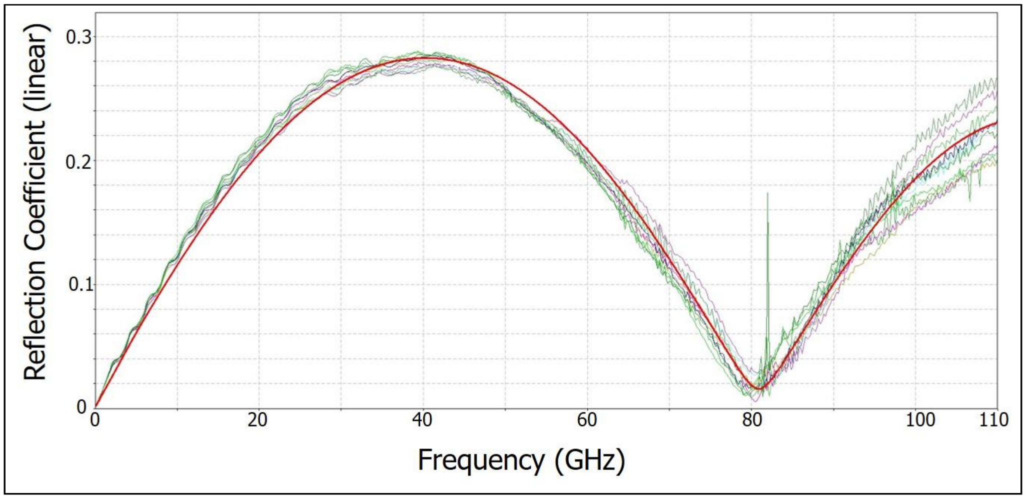


Figure 11. Reflection coefficient of a single 9 µm wide and 1 mm long stripline coupled to an adjacent line 9 µm apart and buried in the third level of metal in the ABF interconnect stack. Simulations are shown in thick solid lines and measurements performed on other die are shown in thinner lines. We do not believe that there is any significance to the isolated spike in the measurements at 82 GHz. The hybrid solver was used for the simulations and only the line width was adjusted in the simulation.

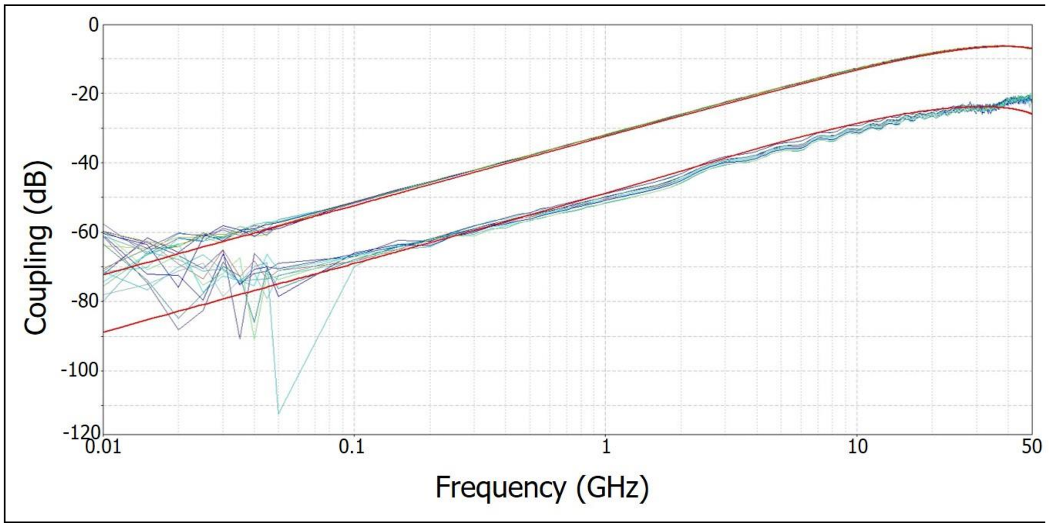


Figure 12. Near-end and far-end coupling between two 9 µm wide and 1 mm long striplines spaced 9 µm apart and buried in the third level of metal in the ABF interconnect stack. Simulations are shown in thick solid lines and measurements performed on other die are shown in thinner lines. The hybrid solver was used for the simulations and only the line width was adjusted in the simulation. The hybrid solver was used for the simulations and only the line width was adjusted in the simulation.

III.3. Eight-Port Coupled Lines on Organic ABF Interposer

Figure 13 shows layouts for two 8-port coupled lines of different lengths. As we did not have access to an 8-port vector network analyzer, we fabricated six versions of each 8-port test structure and connected four of the ports to probe contact pads and terminated the remaining four ports with 50 Ω chip resistors soldered to contact pads on the top of the ABF interposer. These six versions of the 8-port test structures were designed so as to allow us to measure each of the elements of the 8 by 8 scattering-parameter matrix of the coupled lines.

Then we used our four-port analyzer calibrated to a reference plane where the eight-port coupled lines begin and end to measure the various scattering parameters of each of the six test structures. Finally, after compensating for the measured impedances of the chip resistors, we extracted the eight-port scattering parameters of the coupled lines using the Microwave Uncertainty Framework with an algorithm similar to those of [32, 33], except based on a regression fit rather than multiport impedance transformations.

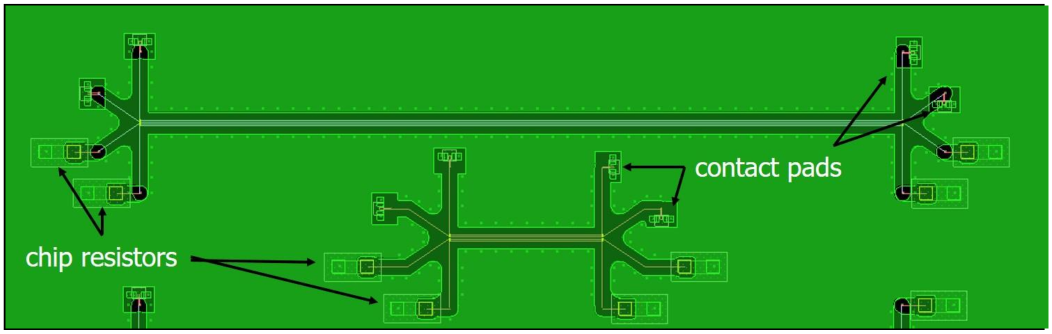


Figure 13. 8-port coupled-line layout.

The manufacturer of the ABF interposer reported much larger variations (as high as ±5 μm) in the line-width measurements performed on the 8-port coupled lines than found on the single-ended and 4-port test structures on the ABF interposer. As a result, we did not expect the level of agreement between the 8-port measurements and simulations as for the other test structures we tested.

In addition, the calibration approach we used was based on the assumption that all of the chip resistors terminating the 8-port coupled lines had an impedance equal to those we measured in a set of chip resistors we characterized separately with a two-port TRL calibration. However, we observed some significant differences between the measurements we performed of the impedances of these chip resistors, possibly due to differences in how each resistor was soldered to the ABF interposer. These measured differences led us to believe that there were also differences in the chip resistors connected to the unmeasured ports of our six 8-port test structures, possibly leading to some additional degradation of the accuracy of the 8-port scattering parameters of the coupled lines we extracted from the measurements.

Williams, Dylan; Chamberlin, Richard; Cheron, Jerome; Chitwood, Sam; Willis, Ken; Butler, Brad; Yazdani, Farhang. "DARPA Organic Interconnect Characterization." Paper presented at DesignCon 2020, Santa Clara, CA, US. January 28, 2020 - January 30, 2020.

Despite the variation in the measured 8-port linewidths and the impedance of the chip resistors on the ABF interposer, we found that our eight-port measurements and simulations agreed quite well. For example, Figs. 14 and 15 compare measurements and simulations of first-neighbor and second-neighbor coupling levels in four 10 mm long 12 µm wide coupled striplines fabricated in the third level of metal in the ABF interconnect stack and separated from each other by 12 µm. The agreement is quite good for first-neighbor coupling and still quite reasonable for the smaller second-neighbor coupling.

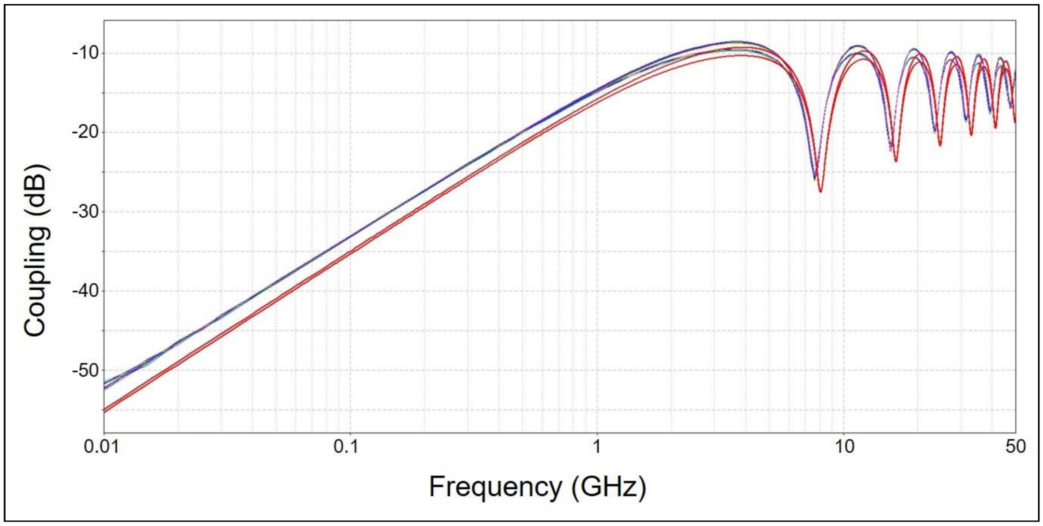


Figure 14. Coupling between first-neighbor 12 µm wide and 10 mm long striplines spaced 12 µm apart and buried in the third level of metal in the ABF interconnect stack. Measurements are shown in thinner lines.

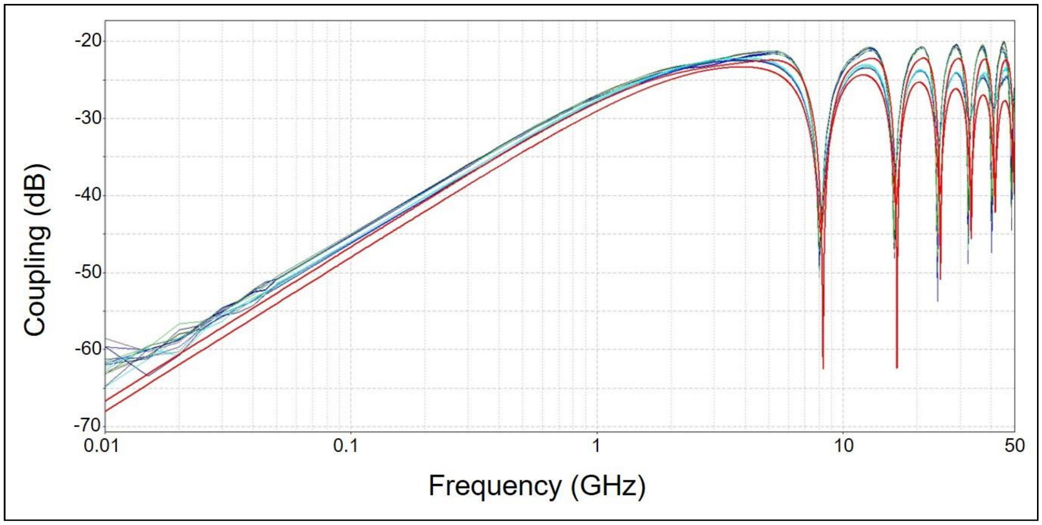


Figure 15. Coupling between second-neighbor 12 µm wide and 10 mm long striplines spaced 12 µm apart and buried in the third level of metal in the ABF interconnect stack. Measurements are shown in the thinner lines.

IV. Conclusion

The measurements and simulations we performed of test structures fabricated in the ABF interconnect stack correlated extremely well, as illustrated by the coupled stripline measurements presented in the last section. We will present more measurements in the conference, including measurements performed on test structures fabricated on the silicon interposer. We will also examine some of the cases we explored on the silicon interposer in which the agreement between the measurements and simulations were not as good as those we observed on the ABF interposer presented here. In general, we found that the slower but more accurate Sigrity full-wave solver was able to provide results that agree closely with the measurements in these cases.

References

[1] R. Mahajan *et al.*, "Embedded Multi-Die Interconnect Bridge (EMIB) – A Localized, High Density Multi-Chip Packaging (MCP) Interconnect," *IEEE Transactions on Components, Packaging and Manufacturing Technology,* pp. 1-1, 2019, doi: 10.1109/TCPMT.2019.2942708.

[2] J. Hruska. "Intel’s New Omni-Directional Interconnect Combines EMIB, Foveros." ExtremeTech. https://www.extremetech.com/computing/294659-intels-new-omni-directional-interconnect-combines-emib-foveros (accessed 2019).

[3] H. Oi, "Evaluation of Ryzen 5 and Core i7 Processors with SPEC CPU 2017," in *2019 IEEE International Systems Conference (SysCon)*, 8-11 April 2019 2019, pp. 1-6, doi: 10.1109/SYSCON.2019.8836790.

[4] H. Oi, "Energy Efficiency Study of Ryzen Microprocessor," in *SoutheastCon 2018*, 19-22 April 2018 2018, pp. 1-5, doi: 10.1109/SECON.2018.8478962.

[5] D. S. Green, C. L. Dohrman, J. Demmin, Y. Zheng, and T. Chang, "A Revolution on the Horizon from DARPA: Heterogeneous Integration for Revolutionary Microwave\/Millimeter-Wave Circuits at DARPA: Progress and Future Directions," *IEEE Microwave Magazine,* vol. 18, no. 2, pp. 44-59, 2017, doi: 10.1109/MMM.2016.2635811.

[6] D. S. Green, C. L. Dohrman, J. Demmin, and T. Chang, "Path to 3D heterogeneous integration," in *2015 International 3D Systems Integration Conference (3DIC)*, 31 Aug.-2 Sept. 2015 2015, pp. FS7.1-FS7.3, doi: 10.1109/3DIC.2015.7334469.

[7] A. Gutierrez-Aitken *et al.*, "A Meeting of Materials: Integrating Diverse Semiconductor Technologies for Improved Performance at Lower Cost," *IEEE Microwave Magazine,* vol. 18, no. 2, pp. 60-73, 2017, doi: 10.1109/MMM.2016.2635838.

[8] A. Gutierrez-Aitken *et al.*, "Diverse Accessible Heterogeneous Integration (DAHI) at Northrop Grumman Aerospace Systems (NGAS)," in *2014 IEEE Compound Semiconductor Integrated Circuit Symposium (CSICS)*, 19-22 Oct. 2014 2014, pp. 1-4, doi: 10.1109/CSICS.2014.6978550.

[9] S. Raman, C. L. Dohrman, and T. Chang, "The DARPA Diverse Accessible Heterogeneous Integration (DAHI) program: Convergence of compound semiconductor devices and silicon-enabled architectures," in *2012 IEEE*

International Symposium on Radio-Frequency Integration Technology (RFIT), 21-23 Nov. 2012 2012, pp. 1-6, doi: 10.1109/RFIT.2012.6401596.

[10] K. K. Samanta, "Pushing the Envelope for Heterogeneity: Multilayer and 3-D Heterogeneous Integrations for Next Generation Millimeter- and Submillimeter-Wave Circuits and Systems," *IEEE Microwave Magazine,* vol. 18, no. 2, pp. 28-43, 2017, doi: 10.1109/MMM.2016.2635858.

[11] R. A. Chamberlin and D. F. Williams, "Measurement and Modeling of Heterogeneous Chip-Scale Interconnections," *IEEE Transactions on Microwave Theory and Techniques,* vol. 66, no. 12, pp. 5358-5364, 2018, doi: 10.1109/TMTT.2018.2873333.

[12] A. Olofsson, D. S. Green, and J. Demmin, "Enabling High-Performance Heterogeneous Integration via Interface Standards, IP Reuse, and Modular Design," *International Symposium on Microelectronics,* vol. 2018, no. 1, pp. 000246-000251, 2018, doi: 10.4071/2380-4505-2018.1.000246.

[13] A. Olofsson. "Common Heterogeneous Integration and IP Reuse Strategies (CHIPS)." DARPA. https://www.darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies (accessed 2019).

[14] "Ajinomoto Build-up Film." The Ajinomoto Group. https://www.ajinomoto.com/en/rd/our_innovation/abf/ (accessed 2019).

[15] "Photosensitive Liquid Solder Resist -SR7300 Series." Hitachi Chemical. http://www.hitachi-chem.co.jp/english/products/pm/022.html (accessed 2019).

[16] R. B. Marks and D. F. Williams, "Verification of Commercial Probe-Tip Calibrations," *Automatic RF Techniques Group Conference Digest,* vol. 42, pp. 37-44, 12/1993 1993.

[17] D. F. Williams and R. B. Marks, "Calibrating On-Wafer Probes to the Probe Tips," *Automatic RF Techniques Group Conference Digest,* vol. 40, pp. 136-143, 12/1992 1992.

[18] D. F. Williams, R. B. Marks, and A. Davidson, "Comparison of On-Wafer Calibrations," in *Automatic RF Techniques Group Conference Digest*, December 1991, vol. 38, pp. 68-81.

[19] R. B. Marks and D. F. Williams, "A general waveguide circuit theory," *J.Res.Nat.Instit.Standards and Technol.,* vol. 97, no. 5, pp. 533-562, September 1992.

[20] D. F. Williams, B. Alpert, U. Arz, D. K. Walker, and H. Grabinski, "Causal characteristic impedance of planar transmission lines," *IEEE Trans Adv.Packaging,* vol. 26, no. 2, pp. 165-171, 5/1/2003 2003.

[21] D. F. Williams and B. Alpert, "Causality and waveguide circuit theory," *IEEE Trans.Microw.Theory Techn.,* vol. 49, no. 4, pp. 615-623, 4/1/2001 2001.

[22] D. F. Williams and B. K. Alpert, "Characteristic impedance, power, and causality," *Microwave and Guided Wave Letters, IEEE,* vol. 9, no. 5, pp. 181-182, 5/1999 1999.

[23] D. Williams, "Traveling Waves and Power Waves: Building a Solid Foundation for Microwave Circuit Theory," *IEEE Microwave Magazine,* vol. 14, no. 7, pp. 38-45, 2013, doi: 10.1109/MMM.2013.2279494.

[24] D. F. Williams *et al.*, "Calibrations for Millimeter-Wave Silicon Transistor Characterization," *IEEE Transactions on Microwave Theory and Techniques,* vol. 62, no. 3, pp. 658-668, 2014, doi: 10.1109/TMTT.2014.2300839.

[25] D. F. Williams, A. C. Young, and M. Urteaga, "A Prescription for Sub-Millimeter-Wave Transistor Characterization," *IEEE Trans.THz Sci.Technol.,* pp. 433-439, July 2013.

[26] D. F. Williams and R. B. Marks, "Transmission Line Capacitance Measurement," *IEEE Microwave and Guided Wave Letters,* vol. 1, no. 9, pp. 243-245, September 1991.

[27] R. B. Marks and D. F. Williams, "Characteristic Impedance Determination using Propagation Constant Measurement," *IEEE Microw.and Guided Wave Letters,* vol. 1, no. 6, pp. 141-143, June 1991.

[28] W. Heinrich, "Full-wave analysis of conductor losses on MMIC transmission lines," *IEEE Transactions on Microwave Theory and Techniques,* vol. 38, no. 10, pp. 1468-1472, 1990, doi: 10.1109/22.58687.

[29] A. R. Djordjevic, R. M. Biljie, V. D. Likar-Smiljanic, and T. K. Sarkar, "Wideband frequency-domain characterization of FR-4 and time-domain causality," *IEEE Transactions on Electromagnetic Compatibility,* vol. 43, no. 4, pp. 662-667, 2001, doi: 10.1109/15.974647.

[30] C. Svensson and G. H. Dermer, "Time domain modeling of lossy interconnects," *IEEE Transactions on Advanced Packaging,* vol. 24, no. 2, pp. 191-196, 2001, doi: 10.1109/6040.928754.

[31] *NIST Microwave Uncertainty Framework*. (2011). National Institute of Standards and Technology, http://www.nist.gov/ctl/rf-technology/related-software.cfm. [Online]. Available: http://www.nist.gov/ctl/rf-technology/related-software.cfm

[32] R. A. Speciale, "A Generalization of the TSD Network-Analyzer Calibration Procedure, Covering n-Port Scattering-Parameter Measurements, Affected by Leakage Errors," *IEEE Trans.Microw.Theory Techn.,* vol. 25, no. 12, pp. 1100-1115, 12/1977 1977.

[33] D. K. Walker, D. Williams, A. Padilla, and U. Arz, "Four-Port Microwave Measurement System Speeds On-Wafer Calibration and Test," *Microwave Journal,* 2001.

Extending NIST's CAVP Testing of Cryptographic Hash Function Implementations

Nicky Mouha and Christopher Celi

National Institute of Standards and Technology, Gaithersburg, MD, USA
`nicky@mouha.be, christopher.celi@nist.gov`

Abstract. This paper describes a vulnerability in Apple's CoreCrypto library, which affects 11 out of the 12 implemented hash functions: every implemented hash function except MD2 (Message Digest 2), as well as several higher-level operations such as the Hash-based Message Authentication Code (HMAC) and the Ed25519 signature scheme. The vulnerability is present in each of Apple's CoreCrypto libraries that are currently validated under FIPS 140-2 (Federal Information Processing Standard). For inputs of about 2^{32} bytes (4 GiB) or more, the implementations do not produce the correct output, but instead enter into an infinite loop. The vulnerability shows a limitation in the Cryptographic Algorithm Validation Program (CAVP) of the National Institute of Standards and Technology (NIST), which currently does not perform tests on hash functions for inputs larger than 65 535 bits. To overcome this limitation of NIST's CAVP, we introduce a new test type called the Large Data Test (LDT). The LDT detects vulnerabilities similar to that in CoreCrypto in implementations submitted for validation under FIPS 140-2.

Keywords: CVE-2019-8741, FIPS, CAVP, ACVP, Apple, CoreCrypto, hash function, vulnerability.

1 Introduction

The security of cryptography in practice relies not only on the resistance of the algorithms against cryptanalytical attacks, but also on the correctness and robustness of their implementations. Software implementations are vulnerable to software faults, also known as bugs.

A (cryptographic) hash function turns a message of a variable length into an output of a fixed length, often called a message digest, or digest. This fixed-length output can then serve as a "fingerprint" for the message, in the sense that it should be computationally infeasible to construct two messages that result in the same digest. Hash functions are crucial to the security of many higher-level cryptographic algorithms and protocols.

In the context of digital signature schemes, hash functions are used to ensure that only the given message and the corresponding signature (along with the public key) passes the signature verification process. Digital signatures provide authentication in a similar manner to signatures in the real world. For example, a web browser can verify a package that is downloaded comes from a specific

Mouha, Nicky; Celi, Christopher. "Extending NIST's CAVP Testing of Cryptographic Hash Function Implementations." Paper presented at CT-RSA 2020: Cryptographers' Track at the RSA Conference, San Francisco, CA, US. February 24, 2020 - February 28, 2020.

website by verifying the signature that was provided with the download using the known, trusted public key of the website. As a part of this verification process, the browser hashes the downloaded data so that the fixed-length digest can stand in place of the large variable-length data in the digital signature scheme.

A recent study by Mouha et al. [12] of the National Institute of Standards and Technology (NIST) SHA-3 (Secure Hash Algorithm) competition found that about half of the implementations submitted to the SHA-3 competition contained bugs, including two out of the five finalists. It appears that cryptographic algorithms can be difficult to implement, given that even the designers of the algorithm can have trouble to develop a correct implementation. Furthermore, even for a secure and well-designed cryptographic algorithm, bugs can be particularly severe with respect to the cryptographic properties of the algorithm's implementation.

For example, in the case of all submitted implementations of the BLAKE [4] algorithm to the SHA-3 competition, given one message and its corresponding hash function output, it is easy to construct another message that produces the same hash value. This "second preimage attack" is not due to a weakness in the BLAKE algorithm specification, but due to an implementation bug that remained undiscovered for seven years.

In [12], Mouha et al. did not find any bugs in the submission packages of Keccak [6], the hash function algorithm that won the SHA-3 competition and that is now standardized in Federal Information Processing Standard (FIPS) 202 [17]. In this paper, we explore whether implementations of hash functions that are standardized by NIST and currently used in commercial products may also contain bugs. Furthermore, we investigate how these bugs can impact more complex cryptographic operations such as digital signature schemes.

2 Testing within NIST's CAVP

NIST maintains the Cryptographic Algorithm Validation Program (CAVP), which provides validation testing for the NIST-recommended cryptographic algorithms. The CAVP is a prerequisite for validating cryptographic implementations according to FIPS 140-2 under the Cryptographic Module Validation Program (CMVP). Since the Federal Information Security Management Act (FISMA) of 2002, U.S. Federal Agencies no longer have a statutory provision to waive FIPS 140-2. This means that commercial vendors must validate their cryptographic implementations, also known as modules, according to CAVP/CMVP before they can be deployed by U.S. Federal Agencies.

The CAVP testing methodology is derived directly from the algorithm specification, independent of the actual code that a vendor's implementation uses. Therefore, it is realistic to expect three main limitations of the CAVP:

1. The CAVP does not require that the internals of an implementation are known in order to generate tests, and is therefore restricted to black-box testing. For many widely-used cryptographic libraries, however, the software

is either open source or available on the vendor's website, which may be used to reveal additional bugs through static analysis (including checking software coding standards), or white-box testing.

2. The CAVP tests only the capabilities of the implementation that are declared by the vendor. For example, a hash function implementation may declare that it can only process messages up to 65 535 bits, corresponding to the largest test vectors currently in the CAVP, even though it may encounter much larger inputs under typical use. When NIST introduces tests for larger inputs, it is therefore the vendor's responsibility to declare whether or not their implementation supports such inputs. However, it is in the vendor's interest to avoid bugs and therefore declare the capabilities of the implementations as broadly as possible.
3. The CAVP focuses mostly on the correct processing of valid inputs (positive testing), rather than the rejection of invalid inputs (negative testing). Because of the nature of black-box testing, the CAVP provides test vector data to the implementation. A developer of the module must program a test harness to submit this data to the interfaces of the cryptographic library itself and collect the output to send back to the CAVP. As the test harness is outside the bounds of the CAVP, it is difficult to know from a validation perspective whether invalid inputs are handled by the module, or by the test harness. There are a few notable exceptions to this, such as the CAVP tests for digital signature schemes that test whether the implementation can recognize valid versus invalid signatures.[1]

Furthermore, the focus of most cryptographic algorithm testing is on correctness towards common cases within the specification. This may leave cryptographic algorithms vulnerable to malicious inputs that manifest themselves very rarely under random testing. Notable examples exploit bugs in modular arithmetic [7], incorrect group order validation [21], or improper primality testing [1] to result in full or partial key recovery attacks on OpenSSL and other implementations. These examples show the importance to consider not just random but also "rare" and "unusual" inputs for cryptographic implementations, as they may lead to catastrophic security failures.

In spite of these limitations, the CAVP can be highly effective at detecting many types of bugs. This is because the CAVP test design is aware of the internals of "typical" implementations of cryptographic algorithms. The focus of the CAVP is not just conformance testing but also regression testing, as the CAVP test design is also aware of how changes to the implementations may lead to certain bugs. To see this, we now explain how the CAVP tests are generated.

The two test types in the CAVP are the Algorithm Functional Test (AFT), and the Monte Carlo Test (MCT). They were introduced in 1977 by the National Bureau of Standards (NBS), the former name of NIST, in the (now-withdrawn) Special Publication (SP) 500-20 [13] to test the Data Encryption

[1] For the signature verification operation, the CAVP also includes some invalid padding tests.

Standard (DES). In this standard, static AFTs known as Known Answer Tests (KATs) were provided in order to “fully exercise the non-linear substitution tables” (S-boxes), whereas MCTs contained “pseudorandom data to verify that the device has not been designed just to pass the [fixed] test set.” Additionally, the large amount of data of the MCT was intended to detect whether it can “cause the device to hang or otherwise malfunction,” for example due to a memory leak [8] in present-day implementations. The spirit and design of these tests was carried over to other algorithms such as the Advanced Encryption Standard (AES) in FIPS 197 [14] and hash functions.

This paper focuses on testing for hash functions within the CAVP at NIST. FIPS 180-4 [16] standardizes the hash functions SHA-1, SHA-224, SHA-256, SHA-384, SHA-512, SHA-512/224, and SHA-512/256. As these hash functions closely resemble each other, they are considered functionally equivalent for the purpose of this document. Testing for SHA-3 was added after the publication of FIPS 202 [16], and with the exception of the SHAKE extendable-output functions (XOFs), mimics the testing done for the FIPS 180-4 hash functions. As with the other CAVP tests, the Secure Hash Algorithm Validation System (SHAVS) [5] specifies both AFTs and MCTs.

Testing by the CAVP was done for many years using the Cryptographic Algorithm Validation System (CAVS) tool. An implementation under test (IUT) is accompanied with a declaration to the CAVS tool of which digest sizes it supports along with a couple of other properties such as whether or not it can hash an empty message, whether or not it can hash incomplete bytes (i.e. a 7-bit message), and the maximum message size. The maximum message size allowed by the tool is 65 535 bits.

As of 2019, the CAVP is undergoing a transition to use the Automated Cryptographic Validation Protocol (ACVP) to enable the generation and validation of standardized algorithm test vectors. This involves a shift of generating and validating tests at remote, approved laboratories, to performing these actions on NIST-hosted servers. The concept of first-party testing is introduced to allow vendors to test and validate their implementations without laboratories as intermediaries. This combined with hosting a demo server (a sandbox environment for algorithm testing), allows vendors to incorporate continuous testing of crypto implementations in their development process. The ACVP thereby significantly speeds up testing and validation.

The ACVP uses a JSON (JavaScript Object Notation) format to specify the test cases. The client to the NIST ACVP servers would then correspond to the test harness in the previous CAVS model, and is responsible for communicating with the server and exercising the proper interfaces on the module. In the JSON examples below, some of the original content has been trimmed for readability. For more information on the protocol itself, as well as the complete examples, we refer to the GitHub repository of the ACVP [11].

2.1 Algorithm Functional Test (AFT)

AFTs take a single message as input, and verify the correctness of the corresponding output. A JSON file is sent from the server to the client, which usually provides inputs to a cryptographic algorithm, and is very simple for an individual test case:

```
{
 "msg": "BCE7",
 "len": 16
}
```

where "msg" corresponds to the message represented as hexadecimal, and "len" corresponds to the length in bits of the message. The messages have fixed values that have been drawn uniformly at random from the space of messages of a certain bit length, ranging from the client's specified minimum to their specified maximum or 65 535, whichever comes first. The expected response to this test case is another simple JSON object:

```
{
 "md": "1FA29E9B23060562F9370453EF817E18C56AE844E5B85F2ED34B4B38"
}
```

where "md" corresponds to the message digest. The hash function in this example is SHA-224.

AFTs can vary in length from byte-oriented messages (i.e., "len" is a multiple of 8) or bit-oriented messages (with any bit lengths). This allows implementations to specify their properties to the CAVP to receive appropriate test cases.

These tests are intended to provide assurance that an implementation can handle messages of various sizes. However, the assurance that the AFTs currently offer may be limited, as they may not test more than one message of any specific bit length.

2.2 Monte Carlo Test (MCT)

MCTs, on the other hand, construct a chain of hash outputs by combining the previous three hash outputs into a single message, and use it to produce the next hash output. Each chain consists of 1000 iterations, and returns the hash output that is obtained at the end. This whole process is repeated 100 times with the original message replaced by the latest hash output.

The initial condition for an MCT is as follows:

```
{
 "msg": "B4FCB616B3A4A7C9E6AF1D836CF1576709A67F16141217B827E52611",
 "len": 224
}
```

where `"msg"` becomes the `seed` in the pseudocode of the MCT, which is given in Alg. 1. The `seed` is not fixed, but is drawn uniformly at random for every invocation of the test.

Algorithm 1 The Monte Carlo Test (MCT) for hash functions

```
Require: seed (random string of same length as hash output)
  for $i = 1$ to 100 do
      MD[0] = MD[1] = MD[2] = seed;
      for $j = 3$ to 1002 do
          Msg[$j$] = MD[$j - 3$] || MD[$j - 2$] || MD[$j - 1$];
          MD[$j$] = Hash(Msg[$j$]);
      end for
      seed = MD[1002];
      Output seed;
  end for
```

The response is an array of 100 hash outputs as follows:

```
{
"resultsArray": [
 {
  "md": "7B893BC7322AA6578A2EC565593B86776FB8376AC16B0A354E6DA016"
 },
 {
  "md": "4BCB655F36D976ADAAE620B485DA7FD8ED321E0BF060E0FE2B5F9AFE"
 },
 {
  "md": "57AA388954B3D52645BFAC69E87F48B3D57A86CF385F38A2549FE957"
 }
]
}
```

shortened to only three outputs for brevity, and again using the SHA-224 hash function in the example. The CAVP makes an implicit assumption here that the client's implementation can handle a message that is three times the size of the hash output.

These tests are intended to provide assurance that an implementation is correct for valid inputs over thousands of iterations. However, the assurance that the MCTs currently offer may be limited, as the bit lengths of the messages do not vary between test cases. Furthermore, as this bit length is three times the digest size, the MCTs only cover a negligibly small percentage of the total input space of the given bit length.

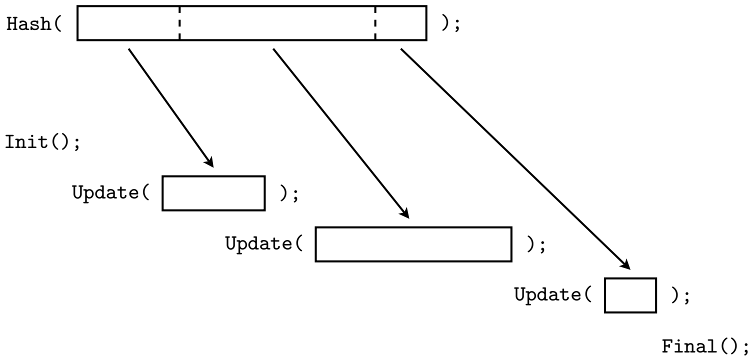


Fig. 1. Hash functions are commonly implemented using a `Hash` interface that takes a variable-length message, and returns a fixed-length output. It is common to also have an `Init-Update-Final` interface, which can be convenient to process large messages on the fly.

3 Common Hashing Interfaces

Although not mentioned in the NIST hash function standards [16, 17], many cryptographic implementations have at least two distinct functional interfaces for hash operations, as shown in Fig. 1. One of the two interfaces, or both interfaces, may be available to a consumer of the module or to higher-level algorithms within the module. The first is an `Init-Update-Final` interface. This structure allows implementations to constantly stream smaller chunks of data into `Update()` repeatedly, rather than keep the message as a single large chunk. Perhaps the entire message is not available at once, or perhaps there is a limit to the capacity of a single `Update()` call.

The other interface is a more intuitive `Hash()` call that expects the whole message up front. This is different from the previous interface and the same module could potentially behave differently under these two interfaces [12].

In practice, the `Init-Update-Final` interface can be convenient to hash the concatenation of various elements. For example, the American National Standards Institute (ANSI) X9.63 Key Derivation Function (KDF) [2] computes the hash of a secret value Z, a counter, and an optional `SharedInfo` string that is shared between two entities. This hash can be computed using one `Init()` call, followed by an `Update()` call to process Z, another `Update()` call for the counter, and then an optional third `Update()` call for `SharedInfo`. The `Final()` call can then be used to compute the hash function output.

To hash the contents of a file, there are two approaches that are commonly encountered in practice. One approach is to loop through the contents of the file (e.g., using `fread()` in C), and process each chunk using a call to `Update()`. Another common approach is to map the file to the virtual address space (e.g., using `mmap()` in C), and then compute the hash by calling `Hash()`. This second approach must be used when the interface requires the data to be located in memory. For example, the interface of the Ed25519 signature scheme in Apple's CoreCrypto requires a pointer for the data to be hashed, therefore if an application wants to compute (or verify) a digital signature on a file (e.g., containing a large software update), it must first use `mmap()` to map this file into memory.

4 Vulnerability in Apple's CoreCrypto Library

We show how adding test cases beyond the current coverage of the CAVP can reveal previously undiscovered bugs in cryptographic implementations.

First, we look the SHAVS document [5], which states that:

> *"While the specification for SHA specifies that messages up to at least $2^{64} - 1$ bits are possible, these tests only test messages up to a limited size of approximately 100,000 bits. This is adequate for detecting algorithmic and implementation errors."*

In contrast, the SHA-3 Competition Test Suite [15] also contains an "Extremely Long Message Test," which contains a message of 2^{33} bits (1 GiB), with the intention of checking whether messages of more than 2^{32} bits were processed correctly. This test from the SHA-3 competition is not adopted by the CAVP however. We now explain how adding a similar test for large messages reveals a bug in the widely-used Apple CoreCrypto library.

Apple makes the source code of its CoreCrypto library publicly available [3] to allow for "verification of its security characteristics and correct functioning."[2] The CoreCrypto library provides low-level cryptographic primitives that are fundamental to the security of Apple's products, and is currently deployed in iPhone, iPad, and Mac devices worldwide. The library has also undergone rigorous testing, and is currently present in 20 FIPS 140-2-validated modules.

In the latest CoreCrypto library, the bug is present in the `ccdigest_update.c` file, which is located in the `ccdigest/src` subdirectory. This code is shared by all implemented hash functions except for MD2. The full code of the function is given in App. A. All the implemented hash functions are iterated hash functions, which means that an underlying compression function processes the message in multiples of a block size that is specific to the algorithm. Part of the code to process message in multiples of the block size is as follows:

```
//low-end processors are slow on division
if (di->block_size == 1<<6 ){ //sha256
  nblocks = len >> 6;
  nbytes = len & 0xFFFFffC0;
} else if(di->block_size == 1<<7 ){ //sha512
```

[2] We refer to the latest CoreCrypto that is available online at the time of writing (November 25, 2019). It does not appear to have a version number, but can be identified by the year 2018 in the copyright notice.

```
  nblocks = len >> 7;
  nbytes = len & 0xFFFFff80;
} else {
  nblocks = len / di->block_size;
  nbytes = nblocks * di->block_size;
}
```

In this code, the variables `len`, `nblocks`, and `nbytes` are declared as `size_t`, which corresponds to a 64-bit unsigned integer on a 64-bit architecture. The `len` variable is the length of the message in bytes. In case `len` is less than 2^{32}, the value of `nblocks` is the number of complete blocks to be hashed: `len` divided by the block size (in bytes), whereas `nbytes` is the number of bytes of these complete blocks.

However, for block sizes of 64 or 128 bytes (i.e., when `di->block_size` is `1<<6` or `1<<7`), the calculation of `nbytes` contains a bug: the four highest bytes of `size_t` are incorrectly set to zero by the bitwise AND (`&`) operation. Consequently, when `len` is at least 2^{32} (corresponding to messages of at least 4 GiB), the value of `nbytes` does not contain the correct number of complete blocks. Therefore, later in the code, the statement `len -= nbytes` does not decrement `len` by the correct amount; instead `len` remains 2^{32} or larger. Given that all these statements are contained in a while-loop with condition `len > 0`, the program enters into an infinite loop.

A list of affected hash function implementations is given in Table 1.

Table 1. Hash function implementations in Apple's CoreCrypto library.

Algorithm	Block size (in bytes)	vulnerable
MD2	16	✗
MD4	64	✓
MD5	64	✓
RIPEMD-128	64	✓
RIPEMD-160	64	✓
RIPEMD-256	64	✓
RIPEMD-320	64	✓
SHA-1	64	✓
SHA-224	64	✓
SHA-256	64	✓
SHA-384	128	✓
SHA-512	128	✓

When this code was written, perhaps the assumption was made that `size_t` corresponds to a 32-bit value, in which case the code would have been correct. When `size_t` is 64 bits, however, the integer constant used to perform the AND operation is incorrect.

One way to avoid this type of bug, could be to follow software coding standards, such as the Computer Emergency Response Team (CERT) C Coding

Standard. This standard states in INT17-C: "Define integer constants in an implementation-independent manner" [19], and gives an example that is very similar to the bug in Apple's CoreCrypto library. Note that it is possible to avoid masks altogether, by using `nbytes = nblocks << 6` or `nbytes = nblocks << 7` for 64-byte and 128-byte blocks respectively.

4.1 Experimental Verification

We downloaded the latest CoreCrypto library from Apple's website [3], and compiled it using the Xcode IDE (Integrated Development Environment) on macOS 10.14 (Mojave) on a mid 2015 MacBook Pro, as well as using Clang 8 under Ubuntu 14.04 on an Intel Skylake processor. For Linux, the README.md file warns that the Linux Makefile is not up-to-date, therefore we needed to make some minor changes to the Makefile to allow compilation.

Because the bug is due to incorrect C code, we expect that the bug will manifest itself on any 64-bit platform for which the code is compiled. To confirm that the executable is stuck in an infinite loop, we added some source code instrumentation.

In our proof of concept code, we generated an input with a length of 2^{32} bytes. Because the actual value of the input is not relevant for the bug, we arbitrarily set all bits to zero in our experiments. When this input is provided to MD4, MD5, RIPEMD-128, RIPEMD-160, RIPEMD-256, RIPEMD-320, SHA-1, SHA-224, SHA-256, SHA-384, or SHA-512, we verified that the implementation enters into an infinite loop. We mentioned earlier that the MD2 implementation does not share the code of `ccdigest_update.c`, and we also confirmed that the same input does not cause an infinite loop for MD2. This provides experimental confirmation for the results of Table 1.

Then, we looked into higher-level cryptographic operations. We found that the implementation of the ANSI X9.63 KDF is not vulnerable when provided with a secret value Z of length 2^{32} bytes. This is due to a range check in the input length, which is documented by the following source code comment in CoreCrypto: "`ccdigest_update only supports 32bit length`."

However, such a range check is not applied to every hash function calculation, and most other cryptographic algorithms inside Apple's CoreCrypto library that use hash functions are vulnerable. We verified that HMAC enters into an infinite loop for all the vulnerable algorithms in Table 1 when provided with a message of 2^{32} bytes.

For the Ed25519 signature scheme, we found that a message of at least $2^{32} + 64$ bytes is needed to trigger the bug. To explain this, note that the Ed22519 algorithm always prepends some data to the message before computing the hash value using SHA-512. This is implemented in Apple's CoreCrypto using the `Init-Update-Final` interface. When there are 64 bytes already in the buffer, the first 64 bytes of the message are used to complete a 128-byte block, which we recall is the block size for the SHA-512 algorithm. After processing the first 64 bytes of the message, if there are at least 2^{32} bytes or more left, then the bug is triggered. For details, we refer to the full code of the `ccdigest_update()` function in App. A.

We verified that the Ed25519 implementation indeed enters into an infinite loop when a message of $2^{32} + 64$ bytes is digitally signed or verified. Note that in order to trigger the bug in the verification operation, it is not necessary to provide a valid signature. Therefore, even if the private key is stored properly and never used to sign long messages, the verification operation still enters into an infinite loop for an incorrectly-signed message of $2^{32} + 64$ bytes or more. Note that digitally signed messages typically come from untrusted sources, because the concern that a message can be modified by an adversary is typically the reason to apply a digital signature in the first place.

Another cryptographic operation in Apple's CoreCrypto that uses hash functions, is the Secure Remote Password (SRP) protocol. This protocol is run between a client and a server, which can create additional security concerns when communication is done over a network and the adversary controls either the client or the server, and may therefore send malicious inputs. In CoreCrypto's SRP implementation, the username is provided as a null-terminated string.

We verified that when this string contains 2^{32} repetitions of the `'a'` character followed by a null character, then the SRP implementation of both the client and the server enter into an infinite loop. Note that in contrast to the previous examples, the length in this case is not provided by the adversary as a separate parameter, but it is derived inside CoreCrypto using C's `strlen()` function. Therefore, range checking all input length values to CoreCrypto would not have been effective to avoid this attack using a long null-terminated string.

In Sect. 2, we recalled that an input that would "cause the device to hang" was already a concern when the MCT test was introduced for DES in 1977. But an infinite loop is also a security vulnerability, categorized under Common Weakness Enumeration (CWE) 835 [20], where it is also known as a "Loop with Unreachable Exit Condition." More specifically, an adversarially-crafted input that causes an implementation to enter an infinite loop, can lead to a "denial of service" (DoS) attack when it consumes excessive CPU resources.

5 Proposing the Large Data Test (LDT)

In the current CAVP tests, the length of the largest message is 65 535 bits. Such small testing sizes are not realistic towards normal usage. We propose a new Large Data Test (LDT) for the CAVP to provide a greater assurance for the implementations that undergo validation.

The LDT would be a type of AFT, and could be specified similarly to the example in Sect. 2.1. Implementations could specify the size of the largest message size that they can handle, for example on the order of 2 GiB to 8 GiB. The ACVP server can select one of many large supported arbitrary sizes to craft messages. However, a test for such messages may be impractical to communicate natively within the normal JSON structures. To work around this limitation, the LDT employs a simple function to generate the test input, as defined in Alg. 2.

Algorithm 2 The Large Data Test (LDT)

```
Require: Msg (a non-zero number of bytes), fullLength (in bits)
  FullMsg = "";
  for $i = 0$ to ceil( fullLength / bitlength(Msg) ) do
      FullMsg = FullMsg || Msg;
  end for
  FullMsg = truncate(FullMsg, fullLength);
  Output FullMsg;
```

Due to the truncation at the end, it is possible for the LDT to output messages of any number of bits, instead of only multiples of the size of the repeating Msg pattern. The Msg pattern itself needs to be an integer number of bytes, in order to greatly simplify implementations in C-like programming languages. This is, however, not an actual restriction to the messages that can be output. The reason is that any 7-bit repeating pattern (for example) can also be written as a 56-bit (= 7-byte) repeating pattern, where 56 is the least common multiple of 7 and 8 (the number of bits in one byte).

With a generator function defined to expand a short message of a few bytes, into a large message of any arbitrary size, we can define the JSON structure for the LDT as the following:

```
{
  "largeMsg": {
    "content": "D6F7",
    "contentLength": 16,
    "fullLength": 34359738368,
    "expansionTechnique": "repeating"
  }
}
```

We define an "expansionTechnique" to allow extensibility in the future for other methods of producing a message of the proper size. In this example "repeating" corresponds to the repeating nature of Alg. 2.

After the test generates a message of a specific number of bits, this message would then be hashed on the server to produce a single hash output similar to the AFTs. Once the test is sent to the client, this could flush out implementations for faults from long messages that produce incorrect outputs. As hashing is a core operation to many other cryptographic operations, it is important to consider scenarios where an adversary may maliciously generate large inputs.

Note that to unearth the bug in the Apple CoreCrypto library, it is necessary to use either the Hash() interface on a message of 4 GiB or more, or the Init-Update-Final interface where at least one of the Update() calls contains 4 GiB or more. In the latter case, it may be necessary to make the message a few bytes longer, as explained in Sect. 4.1.

Given that the LDT is designed to work with large data, we need to take into consideration that the implementation may run out of memory. When allocating dynamic memory (e.g., using `malloc()` in C) or mapping files to the virtual address space (e.g., using `mmap()` in C) are unsuccessful on the target platform, it may be an option to consider increasing the memory available to the platform or even simulating the environment for the purposes of testing.

6 Discussion

As hash functions are a core primitive within many other cryptographic algorithms, it is critically important to ensure correctness under all valid inputs. Yet the methods with which these algorithms are tested are still based on techniques from 1977. While the original tests are still valid, an automated system allows the CAVP to continually add test types and boost the assurances gained from the program. With a publicly standardized JSON protocol, and open-source test harnesses such as libacvp [9], the CAVP is in a good position to move forward with improved testing techniques. We suggest the LDT as a way to directly improve the assurances gained from the CAVP. Of course, one needs to design, specify, publicly review the tests, etc. before they can be used in a program such as CAVP. Openness and transparency are important for acceptance in this highly sensitive domain.

To test the limits of common variable types such as 32-bit unsigned integers, the LDT would need to be on the order of 2^{32} bytes or 4 GiB. This would be sufficient to detect the CoreCrypto bug, and potentially similar bugs in other cryptographic implementations.

However, an inherent limitation of the CAVP and of software testing in general, is that it is a selection process, where a very small subset is selected from the total number of possible test cases. Therefore, testing is not a method to prove the correctness over all types of inputs for an implementation. As stated by Dijkstra, “Program testing can be used to show the presence of bugs, but never to show their absence!” Indeed, the entire goal of software testing is to determine how to perform this selection process, in order to try to quantify the assurance that we get from testing.

Furthermore, the CAVP only tests the capabilities that are declared by the vendor, and would therefore not detect the bug if it only declares support for short messages. While this is reflected in the final validation certificate the vendor receives, this shows the potential need for a wider amount of negative testing. Negative tests are those that test not only well-defined inputs that may be beyond the advertised capabilities, but also invalid inputs.

We note the potential hazards of exposing multiple entry points to a single set of functionality. As mentioned, hash functions often provide at least two interfaces: an `Init-Update-Final` interface and a `Hash()` interface. Often both are exposed such as within CoreCrypto.

Lastly, it can be interesting to explore the parallels between different levels at which vulnerabilities can be handled, as we now explain.

A security vulnerability report to the vendor can allow for a rapid response to address a vulnerability. The FIPS 140-2 Implementation Guidance (IG) [18] encourages this process by providing the vendors with a "means to quickly fix, test and revalidate a module that is subject to a security-relevant CVE." A CVE (Common Vulnerability and Exposure) is security-relevant if it affects how the module meets the requirements of the FIPS 140-2 standard.

For FIPS 140-2 validated cryptographic modules, publishing a vulnerability with a CVE can accelerate the time for end users to obtain crucial security updates. Yet the very nature of the CVE system is an ad hoc procedure, and there is no mechanism in place to ensure that a vendor has learned from such a vulnerability. A vendor may implement test cases within their own development process to detect similar issues in the future, but this holds a very limited scope. The implementations of other vendors could be susceptible to similar issues, but there may be no incentive to react.

If the CAVP implements tests based on CVEs (e.g., as done by Project Wycheproof [10]), then lessons learned from a CVE are not restricted to a single implementation. The requirement of FIPS validation would then also provide stronger assurances to government and private entities that rely on the program. If a CVE can be detected via existing test types, a static test could be seamlessly included from the NIST server. By using an existing test type, no additional code is needed from a test harness to understand how to process the test. In addition, with the speed of testing under ACVP, it is mutually beneficial to constantly test while developing cryptographic implementations.

7 Conclusion

Apple's CoreCrypto library contains a bug due to the implementation-dependent manner in which integer constants are specified. Due to this bug, the MD4, MD5, and the RIPEMD and SHA family hash function implementations enter into an infinite loop for messages of 4 GiB or larger. The bug affects all implemented hash functions (except MD2), and higher-level operations such as HMAC, Ed25519, and SRP. To detect the bug in NIST's CAVP, we proposed a new Large Data Test (LDT) to calculate the hash value for large messages. We also pointed out that stricter coding standards might be helpful to avoid this type of bug.

Responsible Disclosure. The Apple Product Security team was notified of the vulnerability described in this paper on May 30, 2019, and has since taken steps to address the issue. In a conference call on July 17, 2019, Apple Product Security clarified that they do not object to the publication of the research results presented in this paper. On July 23, 2019, Apple Product Security informed us that they are planning to assign a CVE to this issue. On October 29, 2019, Apple publicly announced CVE-2019-8741 to address the vulnerability described in this paper for macOS Catalina 10.15, tvOS 13, watchOS 6, iOS 13, iTunes 12.10.1 for Windows, and iCloud for Windows 7.14.

Acknowledgments. The authors would like to thank the anonymous reviewers and their NIST colleagues for providing useful comments and suggestions. Special thanks go to Patrick Kamongi, Andrew Regenscheid, Apostol Vassilev, and Jeffrey Marron for their detailed feedback. Certain algorithms and commercial products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the algorithms or products identified are necessarily the best available for the purpose.

References

1. Albrecht, M.R., Massimo, J., Paterson, K.G., Somorovsky, J.: Prime and Prejudice: Primality Testing Under Adversarial Conditions. In: Lie, D., Mannan, M., Backes, M., Wang, X. (eds.) Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018. pp. 281–298. ACM (2018), https://doi.org/10.1145/3243734.3243787
2. American National Standards Institute: Public Key Cryptography for the Financial Services Industry - Key Agreement and Key Transport Using Elliptic Curve Cryptography. ANSI X9.63 (2017), https://webstore.ansi.org/standards/ascx9/ansix9632011r2017
3. Apple: Security - Apple Developer (September 2019), https://developer.apple.com/security/
4. Aumasson, J.P., Henzen, L., Meier, W., Phan, R.C.W.: SHA-3 proposal BLAKE. Submission to the NIST SHA-3 Competition (Round 3) (2010), http://131002.net/blake/blake.pdf
5. Bassham III, L.E., Hall, T.A.: The Secure Hash Algorithm Validation System (SHAVS) (May 2014), https://csrc.nist.gov/CSRC/media/Projects/Cryptographic-Algorithm-Validation-Program/documents/shs/SHAVS.pdf
6. Bertoni, G., Daemen, J., Peeters, M., Van Assche, G.: The Keccak SHA-3 submission. Submission to the NIST SHA-3 Competition (Round 3) (2011), http://keccak.noekeon.org/Keccak-submission-3.pdf
7. Brumley, B.B., Barbosa, M., Page, D., Vercauteren, F.: Practical Realisation and Elimination of an ECC-Related Software Bug Attack. In: Dunkelman, O. (ed.) Topics in Cryptology - CT-RSA 2012 - The Cryptographers' Track at the RSA Conference 2012, San Francisco, CA, USA, February 27 - March 2, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7178, pp. 171–186. Springer (2012), https://doi.org/10.1007/978-3-642-27954-6_11
8. Celi, C.: ACVP Secure Hash Algorithm (SHA) JSON Specification. IETF Internet-Draft (2018), https://usnistgov.github.io/ACVP/artifacts/draft-celi-acvp-sha-00.html
9. Cisco: The libacvp library (September 2019), https://github.com/cisco/libacvp
10. Google: Project Wycheproof tests crypto libraries against known attacks. (September 2019), https://github.com/google/wycheproof
11. Industry Working Group on Automated Cryptographic Algorithm Validation: ACVP (September 2019), https://usnistgov.github.io/ACVP/
12. Mouha, N., Raunak, M.S., Kuhn, D.R., Kacker, R.: Finding Bugs in Cryptographic Hash Function Implementations. IEEE Trans. Reliability **67**(3), 870–884 (2018), https://doi.org/10.1109/TR.2018.2847247
13. National Bureau of Standards: Validating the Correctness of Hardware Implementations of the NBS Data Encryption Standard. NBS Special Publication 500-20 (November 1977), https://doi.org/10.6028/NBS.SP.500-20e1977

14. National Institute of Standards and Technology: Advanced Encryption Standard (AES). NIST Federal Information Processing Standards Publication 197 (November 2001), https://doi.org/10.6028/NIST.FIPS.197
15. National Institute of Standards and Technology: Description of Known Answer Test (KAT) and Monte Carlo Test (MCT) for SHA-3 Candidate Algorithm Submissions (February 2008), https://csrc.nist.gov/CSRC/media/Projects/Hash-Functions/documents/SHA3-KATMCT1.pdf
16. National Institute of Standards and Technology: Secure Hash Standard (SHS). NIST Federal Information Processing Standards Publication 180-4 (August 2015), http://dx.doi.org/10.6028/NIST.FIPS.180-4
17. National Institute of Standards and Technology: SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions. NIST Federal Information Processing Standards Publication 202 (August 2015), https://doi.org/10.6028/NIST.FIPS.202
18. National Institute of Standards and Technology and Canadian Centre for Cyber Security: Implementation Guidance for FIPS 140-2 and the Cryptographic Module Validation Program (August 2019), https://csrc.nist.gov/CSRC/media/Projects/cryptographic-module-validation-program/documents/fips140-2/FIPS1402IG.pdf
19. SEI CERT C Coding Standard: INT17-C. Define integer constants in an implementation-independent manner (September 2019), https://wiki.sei.cmu.edu/confluence/display/c/INT17-C.+Define+integer+constants+in+an+implementation-independent+manner
20. The MITRE Corporation: CWE-835: Loop with Unreachable Exit Condition ('Infinite Loop') (2019), https://cwe.mitre.org/data/definitions/835.html
21. Valenta, L., Adrian, D., Sanso, A., Cohney, S., Fried, J., Hastings, M., Halderman, J.A., Heninger, N.: Measuring small subgroup attacks against Diffie-Hellman. In: 24th Annual Network and Distributed System Security Symposium, NDSS 2017, San Diego, California, USA, February 26 - March 1, 2017. The Internet Society (2017), https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/measuring-small-subgroup-attacks-against-diffie-hellman/

A The `ccdigest_update()` function of Apple's CoreCrypto

Here, we provide the implementation of the `ccdigest_update()` in Apple CoreCrypto, which is made available to the public on Apple's website [3]. For readability, we made minor changes to the indentation, corrected the spelling of the word "division" and expanded the `CC_MEMCPY` macro to `memcpy`.

```
void ccdigest_update(const struct ccdigest_info *di, ccdigest_ctx_t ctx,
                     size_t len, const void *data) {
  const char * data_ptr = data;
  size_t nblocks, nbytes;

  while (len > 0) {
    if (ccdigest_num(di, ctx) == 0 && len > di->block_size) {
      //low-end processors are slow on division
      if (di->block_size == 1<<6 ){ //sha256
        nblocks = len >> 6;
```

Mouha, Nicky; Celi, Christopher. "Extending NIST's CAVP Testing of Cryptographic Hash Function Implementations." Paper presented at CT-RSA 2020: Cryptographers' Track at the RSA Conference, San Francisco, CA, US. February 24, 2020 - February 28, 2020.

```
    nbytes = len & 0xFFFFffC0;
   } else if(di->block_size == 1<<7 ){ //sha512
    nblocks = len >> 7;
    nbytes = len & 0xFFFFff80;
   } else {
    nblocks = len / di->block_size;
    nbytes = nblocks * di->block_size;
   }

   di->compress(ccdigest_state(di, ctx), nblocks, data_ptr);
   len -= nbytes;
   data_ptr += nbytes;
   ccdigest_nbits(di, ctx) += nbytes * 8;
  } else {
   size_t n = di->block_size - ccdigest_num(di, ctx);
   if (len < n)
    n = len;
   memcpy(ccdigest_data(di, ctx) + ccdigest_num(di, ctx), data_ptr, n);
   /* typecast: less than block size, will always fit into an int */
   ccdigest_num(di, ctx) += (unsigned int)n;
   len -= n;
   data_ptr += n;
   if (ccdigest_num(di, ctx) == di->block_size) {
    di->compress(ccdigest_state(di, ctx), 1, ccdigest_data(di, ctx));
    ccdigest_nbits(di, ctx) += ccdigest_num(di, ctx) * 8;
    ccdigest_num(di, ctx) = 0;
   }
  }
 }
}
```

Workshop on Machine Learning for Optical Communication Systems: a summary

Josh Gordon
National Institute of Standards and Technology, Communications Technology Laboratory, 325 Broadway, Boulder, CO 80305
email: josh.gordon@nist.gov

Abdella Battou
National Institute of Standards and Technology, Information Technology Laboratory, 100 Bureau Drive, Gaithersburg, MD 20899
email: abdella.battou@nist.gov

Dan Kilper
University of Arizona, James C. Wyant College of Optical Sciences, 1630 E. University Blvd, Tucson, AZ, 85721
email: dkilper@optics.arizona.edu

Abstract: A summary and overview of a public workshop on machine learning for optical Communication systems held on August 2nd 2019, by the Communications Technology Laboratory at the National Institute of Standards and Technology in Boulder, CO.

1. Introduction

Expected increased demand and functionality on optical networks to address higher speeds, lower latency, and higher reliability poses opportunities for improvement from both hardware and software infrastructure. However, the complexity and analog nature of the optical network poses challenges for software control. As these demands increase along with the implementation of smarter components and subsystem the use of artificial intelligence (AI) and machine learning (ML) to improve network function and automation makes increasing sense. Over recent years many use case scenarios have been introduced for ML and AI in optical communication systems; however, to date agreement on implementing ML on a larger scale across the industry and for specific applications is still lacking and debated. Many factors both technical and non-technical play a role in the implementation of ML in optical communication systems (MLOCS).

In an effort to expand discussion on the topic of machine learning (ML) and artificial intelligence (AI) in optical communication systems a public workshop was organized and held at the Boulder campus of the National Institute of Standards and Technology (NIST) by the Communications Technology Laboratory (CTL). On August 2nd 2019, members of industry, academia and government convened to participate in talks, panel discussions and breakout session discussions addressing various subtopics of ML in optical communication systems. The workshop was designed to provide opportunities for attendee participation where scheduled talks transitioned to panel discussions followed by an afternoon session of open breakout discussions on specific subtopics. The workshop attempted to address a range of topics related to AI/ML in the context of optical communication systems. These included general approaches of ML and application to optical transport systems, possible training scenarios, types of data, and use cases. For a more a detailed summary of the workshop outcomes we refer the reader to the workshop White Paper [1] and workshop website at [2].

2. Talk Summary

Workshop talk topics spanned general ML approaches such as linear regression and neural networks as well as training scenarios. The topic of data for use in ML algorithms was also broached in the context of what data is important and the limiting cases of an abundance of data as well as too small of data sets. Several use cases were discussed ranging in domain from the device level to the end-to-end network.

The vast amounts of data generated by optical network management and control shows much promise as a mechanism to make use of AI/ML for the purpose of improving the optical network in several ways from management to improved transmission. In particular quality of transmission (QoT) estimation and failure management became a focus of discussions at the workshop [3],[4]. ML was discussed as a means of improving fault identification and traffic management within the network layer and the function of components in the physical

Gordon, Josh; Battou, Abdella; Kilper, Dan. "Workshop on Machine Learning for Optical Communication Systems: a summary." Paper presented at The Optical Networking and Communication Conference & Exhibition OFC 2020, San Diego, CA, US. March 08, 2020 - March 12, 2020.

layer. Control of amplifiers, mitigation of system nonlinearities, modulation format adaptation and QoT estimation could be addressed with ML. As in most ML approaches these use cases require training data sets, a topic of which was discussed at length during the workshop and deemed a topic requiring ongoing discussion.

Many models were discussed including those for individual devices, network behavior, physical layer impairments, and traffic prediction. Models based either on analytical methods, numerical methods and live monitoring all were considered. Each with advantages and disadvantages, with a tradeoff of accuracy, speed and reliability (e.g. an analytical model may be faster but not as accurate as a complex numerical model, etc.). The use case of nonlinear impairment compensation was a focus of discussion as it is difficult to compute analytically and is a critical factor in many networks, and thus is a scenario wherein ML may prove useful.

Considering the vast amount of data that could be acquired from optical networks the question arises as to “what data matters?”. This is an ongoing question [5],[6],[7],[8] and is use case dependent. Non-linearities in the physical layer poses some of the more challenging scenarios for ML with use cases ranging from nonlinearity mitigation, optical performance monitoring and modulation format recognition. Waveforms show promise as input data sets across use cases whereas on the output what data is useful depends on the use case under consideration. Standardization of waveforms could possibly provide a means for determining the effectiveness of ML algorithms and approaches.

One consideration discussed during the workshop was the data starved scenario, wherein there is a limited set of data available for training. Although the amounts of data could potentially be vast, constraints on data availability whether due to technical or non-technical reasons often severely limit the size of training sets. For example, network operators often cannot share data due to customer privacy considerations. This situation suggests that understanding the case where there is possibly too little data as a result of use case dependent data accessibility is also important. The use of ML and AI spans across many disciplines, where lessons learned, and approaches appear unique to one discipline may in fact inform new approaches and open doors in another. On this topic, work was presented at the workshop on the use of data sets that are “too small” in the context of deep learning for application to computer vision when limited annotated data are available. In particular, two common approaches widely used were discussed: 1) data augmentation and 2) transfer learning. The use of label preserving transformations for data augmentation is an efficient process for expanding a data set that is too small. Furthermore, it also allows one to specify what invariance should be present in the trained model. In the practice of model definition, transfer learning could be used to further refine a model that was first trained on large data sets then refined further on smaller sets of domain or application data.

3. Breakout Session Discussion

Three subtopics were addressed during breakout sessions. As there is a multitude of possible subtopics possible these were judiciously chosen in particular because they provided coverage of several main areas relevant to ML in optical communication systems and overarching architectures. Furthermore, these subtopics contain use cases that range in complexity from individual devices such as coherent transponders, to core networks, to multi-layer networking. More specifically, breakout session topics examined data set curation and applications related to:

- Reconfigurable optical add drop multiplexers (ROADM) and optical line system layer
- Coherent transponders
- Cross Layer End-to-end networking

With these three subtopics the optical line systems (OLS), including the ROADM nodes, fibers and amplifiers were considered as well as transponders and networking. Potential focus areas related to OLS wherein ML could make an impact were quality of transmission (QoT) estimation, fault identification, and failure prediction. QoT estimation was identified as the most promising for obtaining data sets that would provide widespread use across the community. Three network use cases wherein ML applied to the OLS could be an enabler were disaggregation, network defragmentation, and faster dynamic operation. All three of these use cases entail a more “dynamic” network either through the network accommodating variability due to changes and variations in hardware (disaggregation) or through the requirement of more dynamic control (faster add/drop and switching of optical signals). Optical parameters generated by coherent modems were discussed as a potential source of data that may prove difficult to obtain in real time otherwise. Data available from coherent modems was explored in the context of both directly measured as well as computed parameters and the possible applications of such data. In particular, applications range from predictive assessment of network health, fiber health, network performance estimation, to remote assessment of fiber integrity. The complexity of the network as a whole poses challenges to cross-domain core/metro/cloud/access and cross-layer IP/Optical performance and coordination. ML could play an important role in assessing performance based on metrics from such cross-domain/layer interactions. For example, using ML for error correlation, probability, and prediction and impact of errors based on such interactions between domains and layers. Executing ML in such a scenario also highlights the needs for more extensive training data sets.

3. References

[1] NIST Special Publication 2100 “Summary: Workshop On Machine Learning for Optical Communications Systems”. https://doi.org/10.6028/NIST.SP.2100-XX

[2] https://www.nist.gov/news-events/events/2019/08/machine-learning-optical-communication-systems

[3] F. Musumeci, C. Rottondi, G. Corani, S. Shahkarami, F. Cugini and M. Tornatore, "A Tutorial on Machine Learning for Failure Management in Optical Networks," in Journal of Lightwave Technology, vol. 37, no. 16, pp. 4125-4139, 15 Aug.15, 2019.

[4] F. Musumeci et al., "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, Secondquarter 2019.

[5] D. Zibar et al., Machine learning under the spotlight, Nature Photonics, (11) 749-751, 2017.

[6] J. Mata, I. de Miguel, R. J. Durán, N. Merayo, S. K. Singh, A. Jukan, M. Chamania, Artificial intelligence (AI) methods in optical networks: A comprehensive survey, Optical Switching and Networking, 2018.

[7] F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, M. Tornatore, "An Overview on Application of Machine Learning Techniques in Optical Networks," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383-1408, 2019.

[8] F. N. Khan, Q. Fan, C. Lu and A. P. T. Lau, "An Optical Communication's Perspective on Machine Learning and Its Applications," in JLT, vol. 37, no. 2, pp. 493-516, 2019.

Gordon, Josh; Battou, Abdella; Kilper, Dan. ”Workshop on Machine Learning for Optical Communication Systems: a summary.” Paper presented at The Optical Networking and Communication Conference & Exhibition OFC 2020, San Diego, CA, US. March 08, 2020 - March 12, 2020.

PARTICLE TRACKING OF A COMPLEX MICROSYSTEM IN THREE DIMENSIONS AND SIX DEGREES OF FREEDOM

Craig R. Copeland[1], Craig D. McGray[2], B. Robert Ilic[1,3], Jon Geist[2], and Samuel M. Stavis[1]
[1]Microsystems and Nanotechnology Division,
[2]Quantum Measurement Division,
[3]CNST NanoFab, National Institute of Standards and Technology, Maryland USA

ABSTRACT

We make use of the intrinsic aberrations of an optical microscope to track single particles in three dimensions, and we combine information from multiple particles on a rigid body of a microelectromechanical system to measure its motion in six degrees of freedom. Our tracking method provides an extraordinary amount of information from an ordinary imaging system, revealing unintentional motion of the microsystem due to fabrication tolerance and nanoscale clearance between parts in sliding contact. Our work facilitates quantification and study of the actuation performance and reliability of complex microsystems.

KEYWORDS

3D, 6DOF, MEMS, metrology, microscopy, motion

INTRODUCTION

Mechanical systems are fundamentally important to many modern technologies and have great potential for future development. Complex mechanical systems often use rigid-link mechanisms to transduce an input into an output of forces and motion to reliably perform useful work. Such transduction can involve the transfer of motion, with multiple degrees of freedom, through parts of the system in sliding contact. However, practical limitations of transduction result in an imperfect output of forces and motion, limiting performance and reliability. In particular, coupling interactions between parts in sliding contact such as play, feedback, friction, and wear can affect the system kinematics. Play and feedback can not only reduce the precision of the intentional motion of the system but can also result in unintentional and unpredictable motion of microsystems [1]. Moreover, friction and wear can become problematic at the microscale and nanoscale, due to high surface to volume ratios and lack of lubricating liquid films, motivating efforts to elucidate and mitigate these effects [2-5]. In the case that coupling interactions result in nondeterministic motion at small scales, it is necessary to measure single motion with cycles of the system to understand the true kinematics. In any case, single motion cycles are essential in a variety of applications in which the motion of the system is fundamentally aperiodic, such as positioning, switching, and gripping. However, methods to measure the single motion cycles of mechanical systems – often occurring at length scales much smaller than the parts of the system, ranging from the millimeter scale down to the atomic scale, and occurring in six degrees of freedom – have lagged behind microsystem technologies.

Methods for measuring micromechanical motion are often optical due to compatibility with the operation of the systems. Imaging methods can resolve single cycles of in-plane motion at the nanometer scale but have limits of the magnitude or mode of motion that is measurable [6, 7], specific requirements for system design [8], or the need to fabricate test structures for imaging [9]. Interferometry methods for measuring out-of-plane motion can be precise but have only recently measured in-plane motion [10] and require scanning, impeding characterization of systems with multiple moving parts. These various limitations leave a measurement gap, presenting both opportunities and challenges for the development of complex microsystems.

To bridge this gap, we advance our method of particle tracking [1, 11-14] to provide a new capability to measure microsystem motion. We exploit the intrinsic aberrations of an optical microscope to track single particles in three dimensions on a complex microsystem [15, 16], which moves through a deep and ultrawide field. We introduce an algorithm for position estimation in three dimensions, using light-weighting [17] to robustly fit Gaussian models to images and precisely estimate image shape throughout the focal volume. Linear combinations of Zernike polynomials extend this localization analysis to a widefield calibration. A rigid transform combines information from multiple particles to measure the output motion of the microsystem in six degrees of freedom, enabling tracking of the system as it rotates through a repetitive orbit with nanoscale deviations from coupling interactions. Our new method is broadly applicable, and our measurement results provide new insight into the actuation performance and reliability of complex microsystems.

RESULTS AND DISCUSSION

Materials and methods

Our microscope system has an objective lens with a nominal magnification of 50× and a numerical aperture of 0.55. A piezoelectric actuator translates the lens in the z direction with a resolution of 10 nm, providing reference values of z that we assume are accurate. A light-emitting diode illuminates the sample at a peak wavelength of approximately 625 nm. A complementary metal-oxide-semiconductor camera operating with a global shutter records fluorescence micrographs at a peak wavelength of approximately 645 nm. We calibrate the mean value of image pixel size as 127.34 nm ± 0.03 nm [17]. We report all uncertainties in this study as standard deviations, or we note otherwise. We will present further details and calibration of our microscope system in a future study.

Our test microsystem features a rotational electrostatic actuator coupling through a ratchet mechanism to a ring gear, forming a drive motor that operates in an open loop [15, 16]. The ring gear has 200 teeth that couple to a load gear with 80 teeth and a nominal diameter of 328 µm, as Figure 1a shows. A constellation of fluorescent particles with nominal diameters of 1 µm rides on the load gear, as Figure 1b-c shows. A square wave voltage with an

Copeland, Craig R.; McGray, Craig; Ilic, Robert; Geist, Jon; Stavis, Samuel M. "Particle Tracking of a Complex Microsystem in Three Dimensions and Six Degrees of Freedom." Paper presented at 2020 IEEE 33rd International Conference on Micro Electro Mechanical Systems (MEMS), Vancouver, CA. January 18, 2020 - January 22, 2020.

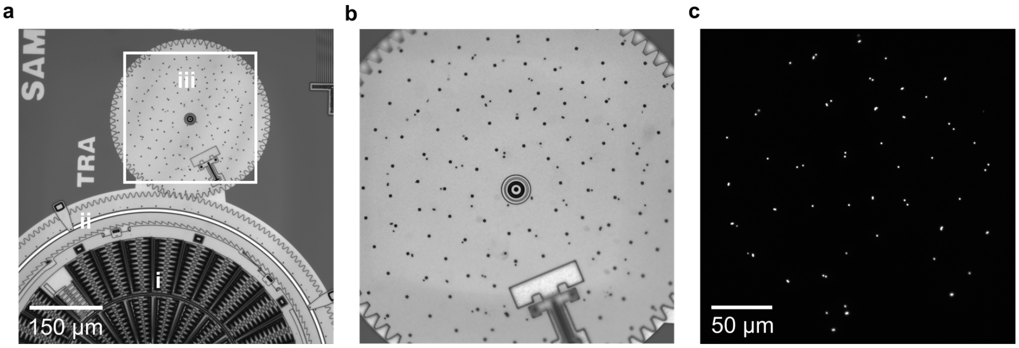


*Figure 1. Experimental overview. (**a**) Brightfield micrograph at inspection magnification showing the drive motor, with (i) a rotational actuator, (ii) a ring gear, and (iii) the load gear. (**b**) Brightfield micrograph at experimental magnification showing the load gear. The smaller dots with random spacing are fluorescent particles and the larger dots in a radial pattern are etch holes for sacrificial release of the structures. (**c**) Fluorescence micrograph at experimental magnification showing a constellation of fluorescent particles on the surface of the load gear.*

amplitude and offset of approximately 7.5 V and a frequency of approximately 90 Hz actuates the motor. Each period incrementally rotates the load gear, which we define as a single motion cycle. After each motion cycle, the microscope system records a fluorescence micrograph of the particles on the quasi-static gear, as Figure 1c shows.

We fit models to data by the method of damped least-squares with uniform weighting, or we note otherwise.

Particle tracking in three dimensions

Intrinsic aberrations of our optical microscope affect the peak amplitudes and shapes of particle images, as Figure 2a shows, enabling localization in three spatial dimensions. We fit a bivariate Gaussian model to the image data by the light-weighting algorithm [17], exploiting the information content of multiple parameters,

$$G_{biv}(x_p, y_p) =$$
$$A \cdot \exp\left(\frac{-1}{2(1-\rho^2)}\left[\frac{(x_p-x)^2}{\sigma_x^2} - 2\rho\frac{(x_p-x)(y_p-y)}{\sigma_x\sigma_y} + \frac{(y_p-y)^2}{\sigma_y^2}\right]\right) + B$$

where (x_p, y_p) is the position of a pixel in a micrograph, (x, y) is the position of the Gaussian peak, A is the peak amplitude, σ_x is the standard deviation in the x direction, σ_y is the standard deviation in the y direction, ρ is the correlation coefficient between the x and y directions, and B is a constant background. Division of ρ by the amplitude parameter A, $\rho_{amp}=\rho/A_n$, where $A_n= A/A_{\rho=\rho^*}$ is the amplitude parameter after normalization to its value in the particular image for which $\rho = \rho^*$, provides a useful z dependence, as Figure 2b shows. This normalization ensures that any differences in fluorescence intensity between the particles for calibration and for experiment do not affect the calibration of z dependence. We set ρ^* to the minimum value of $|\rho|$. The statistical uncertainty of axial localization from empirical polynomial models of z(ρ_{amp}), such as Figure 2b shows, is approximately 50 nm and is due primariliy to the random noise in ρ_{amp}. To accurately calibrate the lateral dependence of the relationship in Figure 2b, we deposit a random array of calibration particles onto a silicon substrate and image the particles through focus, sampling the field to obtain a set of local calibration functions $[z(\rho_{amp})]_{cal}$. The z positions from this set of calibration functions vary significantly across the lateral extent of the field of 260 µm by 260 µm, as Figure 2c shows for a representative value of ρ_{amp} = -0.1.

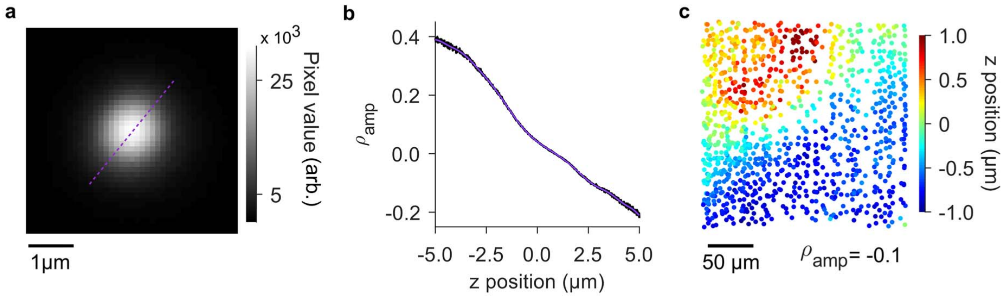


Figure 2: Axial localization. (a) Optical micrograph showing an image of a fluorescent particle with a diameter of approximately 1 µm at a z position of 2 µm above best focus. The dash violet line indicates the major axis. (b) Plot showing a representative dependence of the shape parameter $\rho_{amp}=\rho/A_n$ on z position, where ρ is the correlation coefficient and A_n is the normalized amplitude of a bivariate Gaussian model. The solid violet line shows a polynomial of order 16, which empirically models the z dependence for axial localization. (c) Plot showing the lateral dependence of z position for a common value of the shape parameter ρ_{amp} = -0.1 from a random array of calibration particles that sample the imaging field. Uncertainties in (c) are smaller than the data markers.

Copeland, Craig R.; McGray, Craig; Ilic, Robert; Geist, Jon; Stavis, Samuel M. "Particle Tracking of a Complex Microsystem in Three Dimensions and Six Degrees of Freedom." Paper presented at 2020 IEEE 33rd International Conference on Micro Electro Mechanical Systems (MEMS), Vancouver, CA. January 18, 2020 - January 22, 2020.

To develop our widefield calibration, we model this lateral dependence by linear combinations of Zernike polynomials. The residuals of this model fit to data such as in Figure 2c define systematic errors of axial localization, which have root-mean-square values of approximately 70 nm. These errors are due primarily to variation of image shape at lateral scales of a few micrometers, as Figure 2c shows. We perform analogous calibrations for the axial and lateral dependence of apparent lateral motion. We will elucidate these results, as well as the precision and accuracy of our method, in a future study.

Microsystem tracking in six degrees of freedom

A rigid transform maps particle positions between subsequent micrographs by the iterative closest-point algorithm, yielding the motion of the gear for each cycle. The center of rotation is a natural origin of our extrinsic coordinate system, which we determine as the mean value of all particle positions in each dimension over multiple revolutions of the gear. The residuals of the rigid transforms confirm the accuracy of tracking single particles to within root-mean-square errors of approximately 2 nm in x and y and approximately 80 nm in z. The rigid transform determines three translations Δx, Δy, and Δz, and three rotations in as many degrees of freedom, the intrinsic rotation of the gear γ about the axis of rotation, the nutation β of the axis of rotation with respect to the extrinsic z axis, and the precession α of the axis of rotation about the extrinsic z axis, as Figure 3 shows. Uncertainties of these six degrees of freedom depend on the particular particle constellation and the extrinsic coordinate system [13, 18], as we will describe in a future study.

Our measurements show significant variation in four of the six degrees of freedom due to play in the couplings between the parts of the microsystem, as Figure 3 shows. These results provide new insight into how the parts of complex microsystems move during operation, including both intentional and unintentional motion at nanometer and milliradian scales. Clearance between the meshing teeth of the ring and load gears, as well as variability of the rotational actuator [12], cause variability of the intrinsic rotation γ in Figure 3a. Clearance between the gear and the hub causes variability of the translations Δx and Δy that Figure 3d,e shows. Clearance between the gear and the underlying substrate allows for nutation β of the axis of rotation relative to the extrinsic z axis, as Figure 3c shows. The gear does not exhibit significant translation in the z direction Δz, as Figure 3f shows, which validates the measurement uncertainty for translations in this direction. While the precession of the axis of rotation α varies over a wide range, as Figure 3c shows, the small nutation β that Figure 3b shows causes the rigid transform to be relatively insensitive to this degree of freedom, so that nearly all of the variability is within measurement uncertainty. A different selection of reference frame could improve some of these uncertainties while degrading others. We will analyze these results in more detail in a future study.

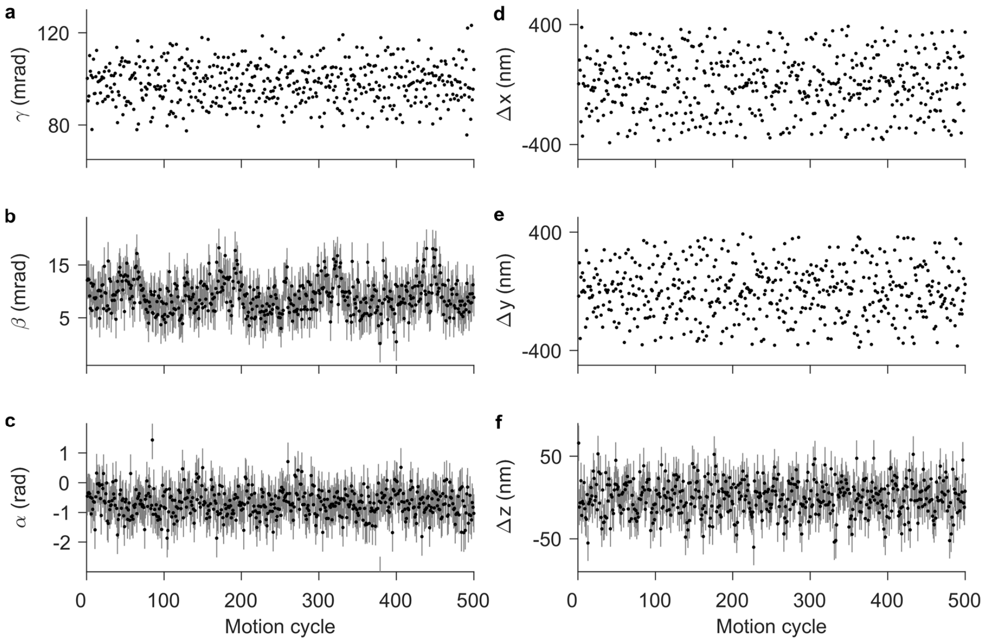


Figure 3: Microsystem tracking. (a-c) Plots showing (a) intrinsic rotations of the gear in three-dimensional space γ , (b) the angle between the axis of rotation and the extrinsic z axis, or nutation β, and (c) the angle between the axis of rotation and the extrinsic x-z plane, or precession α. Variability in (a-c) is due to rotational play in three dimensions. (d-f) Plots showing translation of the gear in the (d) x, (e) y, and (f) z directions. Variability in (d) and (e) indicates translational play, with effective ranges of 394 nm and 387 nm, respectively, while variability in (f) is due to measurement uncertainty. Uncertainties in (a, d, e) are smaller than data markers.

Copeland, Craig R.; McGray, Craig; Ilic, Robert; Geist, Jon; Stavis, Samuel M. "Particle Tracking of a Complex Microsystem in Three Dimensions and Six Degrees of Freedom." Paper presented at 2020 IEEE 33rd International Conference on Micro Electro Mechanical Systems (MEMS), Vancouver, CA. January 18, 2020 - January 22, 2020.

CONCLUSIONS

We introduce a method of particle tracking in three dimensions to measure the motion of a microscale rigid body with six degrees of freedom. Rigid transforms allow the combination of localization data to improve centroid and orientation precision through the central limit theorem [13], now in three dimensions. While such analysis is common in motion tracking of macroscale objects, this is the first such application in particle tracking by optical microscopy. Our new method enables study of a complex microsystem. We find that microfabrication tolerance and nanoscale clearance between parts in sliding contact not only degrade the precision of the intentional motion of the test system, but also result in unintentional motion that is orthogonal to the primary plane of the test system and occurs in all six degrees of freedom. These results provide insight into actuation performance and reliability. While our test system dates back two decades and is commercially available, its ultraplanar fabrication process remains near the state of the art, and we are unaware of any previous measurement that reveals the complex kinematics resulting from nanoscale clearances within such a system. Moreover, commercial microsystems still do not make full use of the complexity that such processes enable, and indeed most microsystems have simpler kinematics than our test system. Advancing practical methods to quantify and specify the motion of complex microsystems will help to fulfill their potential for many exciting applications.

ACKNOWLEDGEMENTS

The authors acknowledge support under the NIST Innovations in Measurement Science program. C.R.C. acknowledges support under the Cooperative Research Agreement between the University of Maryland and the National Institute of Standards and Technology Center for Nanoscale Science and Technology, Award 70NANB10H193, through the University of Maryland.

REFERENCES

[1] C. R. Copeland, C. D. McGray, J. Geist, V. A. Aksyuk, and S. M. Stavis, "Transfer of motion through a microelectromechanical linkage at nanometer and microradian scales," *Microsystems & Nanoengineering,* Article vol. 2, 2016.

[2] I. Rosenhek-Goldian, N. Kampf, A. Yeredor, and J. Klein, "On the question of whether lubricants fluidize in stick–slip friction," *Proceedings of the National Academy of Sciences,* vol. 112, no. 23, pp. 7117-7122, 2015.

[3] A. Vijayasai, G. Sivakumar, G. Ramachandran, C. Anderson, R. Gale, and T. Dallas, "Characterization of a MEMS tribogauge," *Surface and Coatings Technology,* vol. 215, no. 0, pp. 306-311, 2013.

[4] T. D. B. Jacobs and R. W. Carpick, "Nanoscale wear as a stress-assisted chemical reaction," *Nat Nano,* vol. 8, no. 2, pp. 108-112, 2013.

[5] F. W. DelRio, M. P. de Boer, J. A. Knapp, E. David Reedy, P. J. Clews, and M. L. Dunn, "The role of van der Waals forces in adhesion of micromachined surfaces," *Nature Materials,* vol. 4, no. 8, pp. 629-634, 2005.

[6] D. M. Freeman, "Measuring Motions of MEMS," *MRS Bulletin,* vol. 26, no. 04, pp. 305-306, 2001.

[7] *MSA 500 Hardware Manual*. Polytec (Hopkinton, MA).

[8] C. Yamahata, E. Sarajlic, G. J. M. Krijnen, and M. A. M. Gijs, "Subnanometer Translation of Microelectromechanical Systems Measured by Discrete Fourier Analysis of CCD Images," *J. Microelectromech. Syst.,* vol. 19, no. 5, pp. 1273-1275, 2010.

[9] P. Cheng and C. H. Menq, "Real-Time Continuous Image Registration Enabling Ultraprecise 2-D Motion Tracking," (in English), *IEEE Trans. Image Process.,* Article vol. 22, no. 5, pp. 2081-2090, 2013.

[10] C. Rembe, R. Kowarsch, W. Ochs, A. Dräbenstedt, M. Giesen, and M. Winter, "Optical three-dimensional vibrometer microscope with picometer-resolution in x, y, and z," *OPTICE,* vol. 53, no. 3, pp. 034108-034108, 2014.

[11] C. R. Copeland, C. D. McGray, J. Geist, V. A. Aksyuk, and S. M. Stavis, "Characterization of electrothermal actuation with nanometer and microradian precision," in *Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS), 2015 18th International Conference on*, 2015, pp. 792-795

[12] C. R. Copeland, C. D. McGray, J. Geist, and S. M. Stavis, "Particle tracking of microelectromechanical system performance and reliability," *J. Microelectromech. Syst.,* vol. 27, no. 6, pp. 948-950, 2018.

[13] C. McGray, C. R. Copeland, S. M. Stavis, and J. Geist, "Centroid precision and orientation precision of planar localization microscopy," *Journal of Microscopy,* pp. 238-249, 2016.

[14] C. D. McGray *et al.*, "MEMS Kinematics by Super-Resolution Fluorescence Microscopy," *J. Microelectromech. Syst.,* vol. 22, no. 1, pp. 115-123, 2013.

[15] S. M. Barnes, S. L. Miller, M. S. Rodgers, and F. Bitsie, "Torsional ratcheting actuating system," in *2000 International Conference on Modeling and Simulation of Microsystems, Technical Proceedings*, M. Laudon and B. Romanowicz, Eds., 2000, pp. 273-276

[16] D. M. Tanner *et al.*, "Reliability of a MEMS Torsional Ratcheting Actuator," in *39th Annual Proceedings: International Reliability Physics Symposium 2001*, 2001, pp. 81-90

[17] C. R. Copeland *et al.*, "Subnanometer localization accuracy in widefield optical microscopy," *Light: Science & Applications,* vol. 7, 2018.

[18] M. Shah, M. Franaszek, and G. Cheok, "Propagation of Error from Registration Parameters to Transformed Data," *Journal of Research of the National Institute of Standards and Technology,* vol. 121, 2016.

CONTACT

*S.M. Stavis, samuel.stavis@nist.gov

Copeland, Craig R.; McGray, Craig; Ilic, Robert; Geist, Jon; Stavis, Samuel M. "Particle Tracking of a Complex Microsystem in Three Dimensions and Six Degrees of Freedom." Paper presented at 2020 IEEE 33rd International Conference on Micro Electro Mechanical Systems (MEMS), Vancouver, CA. January 18, 2020 - January 22, 2020.

Microring resonator-coupled photoluminescence from silicon W centers

A N Tait[1], S M Buckley[1], J Chiles[1], A N McCaughan[1], S Olson[2], S Papa Rao[2,3], S W Nam[1], R P Mirin[1] and J M Shainline[1]

[1] Applied Physics Division, National Institute of Standards and Technology, Boulder, CO 80305, United States of America
[2] NY CREATES, Albany, NY 12203, United States of America
[3] SUNY Polytechnic Institute, Albany, NY 12203, United States of America

E-mail: alexander.tait@nist.gov

Keywords: silicon light sources, microresonators, photoluminescence, cryogenic photonics, silicon waveguides

Abstract

Silicon defect centers are promising candidates for waveguide-integrated silicon light sources. We demonstrate microresonator- and waveguide-coupled photoluminescence from silicon W centers. Microphotoluminescence measurements indicate wavelengths on-resonance with resonator modes are preferentially coupled to an adjacent waveguide. Quality factors of at least 5,300 are measured, and free spectral ranges closely match expectation. The W center phonon sideband can be used as a spectral diagnostic for a broader range of waveguide-based devices on cryogenic silicon photonic platforms.

1. Introduction

The rapid growth of the silicon photonics industry has the potential to bring photonic manufacturing economies comparable to those of the silicon microelectronics industry. In addition to potentials for large-volume production, silicon photonics opens possibilities for large-scale photonic processing architectures that could not be conceived of in fiber or III-V platforms [1–3]. In all photonic systems, light sources are required. Silicon does not readily emit light at room temperature due to its indirect bandgap. Therefore, most efforts in silicon photonics use external light sources coupled to chip with fiber. The use of external light sources presents critical burdens of fiber packaging, resulting in severe scaling limitations.

Substantial research has been dedicated to developing integrated light sources for silicon photonics [4, 5]. Each approach has advantages and drawbacks. These include rare earth element doping (low brightness), wafer bonding of III-V quantum wells [6] (challenging and expensive processing), epitaxial growth of III-V quantum dots [7] (specialized epitaxy steps), and strain engineering of germanium [8] (fragility and low yield). All of these approaches with the exception of germanium involve the introduction of materials that are incompatible with typical silicon foundries.

Silicon emitters can be realized by creating specific defects in the silicon crystal. Emissive defect centers in silicon have been studied for seven decades, reviewed in [9, 10], but they have not demonstrated utility as sources for silicon photonic circuits in key applications of interest. Their application domains are limited by a requirement of low-temperature operation; however, this is a limitation of little consequence when cryogenic operation is acceptable or required. Defect centers possess several important advantages, most crucial of which is their ease of fabrication: local modifications to the native crystal as opposed to the introduction of new materials and/or structures. If monolithically integrated with waveguides, defect center light sources could provide a compact, inexpensive, scalable solution for superconducting optoelectronic neural networks [2] and optical quantum information systems. In optical quantum information systems, the complex photonic circuitry can operate at room temperature, but then off-chip cryogenic single-photon detectors are needed anyways [11, 12]. The prospect of monolithic sources and single-photon detectors could potentially justify the overhead of cooling the entire quantum system.

In this work, we explore the coupling of W-center photoluminescence to silicon-on-insulator waveguides and microring resonators, a platform shown in figure 1. W centers are trigonal defects that have a

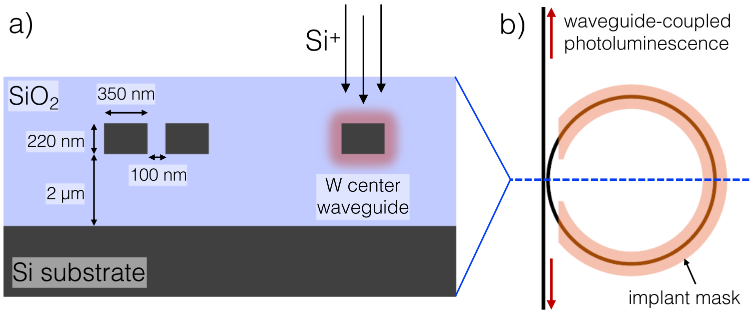


Figure 1. Illustration of platform and devices. (a) Waveguide cross-sections. All waveguides are designed to be 220 nm thick and 350 nm wide such that they are single mode at 1218 nm. As fabricated, waveguide widths were 390 nm. W-center emitters are created by implanting a waveguide with silicon ions and annealing. Evanescent couplers between microrings and the bus waveguide have 100 nm gaps in this work. (b) Top view of microring resonator coupled to a waveguide. Black is silicon; red is the pattern of ion implants defined lithographically; blue, dashed line is the cross-section slice. Implant regions are widened to relax alignment tolerance. Ions that miss the silicon become embedded in the buried oxide and have no effect on the studied devices.

zero-phonon line at 1218 nm and phonon sideband over 1225 nm–1275 nm [13]. They are created by implanting silicon crystal with silicon ions. We report that photoluminescence emitted into resonator modes preferentially couples to a bus waveguide, resulting in spectra with periodic resonances that are characteristic of microrings. The free spectral ranges of different devices closely match with theoretical models, thus allowing extraction of the group velocity parameter with a $R^2 = 0.991$ coefficient of determination. We record resonances with quality factors up to 5,300. Infrared camera and polarization measurements are used to confirm correct operation of the experimental technique for differentiating waveguide-coupled from free-space-coupled photoluminescence.

In prior work, W center waveguide coupling has been achieved in an electrically pumped device whose total power output was measured by an on-chip detector [14]. While the current work aims to support development of electrically pumped devices, micro-photoluminescence (PL) measurements offer additional research opportunities and richer information than electrical devices alone, such as pumping an array of devices or resolving emission spectra. In prior micro-PL measurements of defect centers in silicon-on-insulator, waveguides were not investigated, although several of these devices were on silicon-on-insulator of dimensions that could potentially support waveguides [15–17]. Other studies have shown coupling of various defect centers to suspended microdisk [18] and photonic crystal [19, 20] cavities; however, PL collection has always occurred directly above the structure being pumped, i.e. normal to the surface of the sample.

We monitor waveguide-coupled luminescence unambiguously by collecting from a grating coupler that is spatially offset from the microring being pumped. This circuit is shown in figure 2. The experimental apparatus for collecting offset photoluminescence is similar to those used in prior research on quantum dots and photonic crystals [21–23]. In contrast to those works, both the resonator under study and the offsetting circuitry itself are based on oxide-clad, single-mode silicon rib waveguides of the same type that form the basis of mainstream silicon photonic platforms. Results imply a potential to integrate W centers in more complex photonic integrated circuits and an ability to generalize the experimental technique to the study of a wide range of more complex photonic circuits with embedded emissive defect centers.

2. Layout and fabrication

The substrate used for this study was a 76 mm silicon-on-insulator (SOI) wafer with a 220 nm silicon device layer on a 2 μm buried oxide. Implants were masked by 600 nm thick positive-tone resist patterned by electron-beam lithography. The wafer was then implanted with Si ions at a commercial facility. The silicon etch pattern was masked by an electron-beam-patterned oxide hard mask. The 220 nm thick device layer was fully etched to the buried oxide with a reactive ion etch based on SF_6 and C_4F_8. Finally, the wafer was encapsulated by a 1.5 μm oxide. The wafer was diced into 1 cm × 1 cm die, which were then annealed. Implant and anneal conditions were chosen to yield the optimum W-center brightness as determined in reference [17].

Waveguides were designed to be 350 nm wide such that they are single-mode at 1220 nm. Scanning electron micrographs of the waveguides showed an actual width of 390 nm. Comparisons of insertion loss in single-mode waveguides of different lengths indicated a waveguide loss of 18 dB cm^{-1}, which is significantly

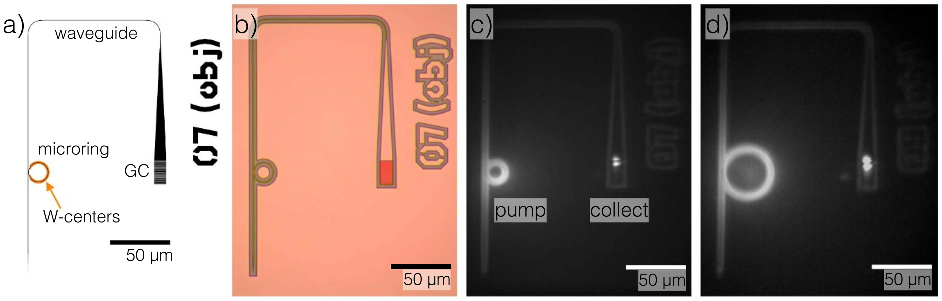


Figure 2. Different views of the circuit for offset collection from a microring resonator (MRR) with radius $r = 8$ μm. (a) Design layout. Silicon is black and the location of W-center implants is orange. GC: grating coupler. (b) Visible microscope image of the fabricated circuit. (c) Infrared image of the circuit when the MRR is optically pumped. Photoluminescence is observed directly from the MRR (free-space-coupled) and also from the offset grating coupler (waveguide-coupled). (d) Infrared image of a different device with larger radius, $r = 20$ μm.

higher than expected due to a non-optimized etch recipe. The peak brightness of the W-center zero phonon line (ZPL) in an unpatterned SOI region was measured to give 4,400 ± 220 counts/s with a 300 μW pump, which agrees with different cooldowns of different samples on the same setup. Microrings were designed with coupling gaps of 100 nm and radii of $r(\mu\text{m}) = [2, 4, 8, 10, 20, 30]$.

We employ a waveguide-based offset collection technique that allows for observation of waveguide-coupled PL, which was introduced in our recent conference proceeding [24]. The silicon photonic circuit, shown in figure 2, spatially separates waveguide-coupled PL from free-space-coupled PL, and the optical apparatus isolates these two collection points. In the circuit, an MRR implanted with W centers is evanescently coupled to a bus waveguide. When the MRR is pumped, some component of the PL is emitted upward into free-space. Another component of the PL is emitted into the resonator modes. Light in the resonator modes couples to the bus waveguide and is then routed to the spatially offset grating coupler (GC). The GC is designed to emit normal to the surface and is located 100 μm away from the center of the MRR.

3. Experimental Apparatus

A simplified experimental setup is shown in figures 3(a) and (b) and in more detail in the Supplementary Information. The primary components are a pump, cryostat, objective, and infrared (IR) spectrometer. The 11 mW laser pump at 635 nm (red) is focused onto the device in the cryostat. Excited electrons undergo an optical transition in the W centers, emitting IR photoluminescence in the infrared (depicted as blue). This PL is directed to a spectrometer and InGaAs camera. PL is emitted from two locations: the device itself and the offset GC. When the PL is imaged with a lens, the two emission sources are spatially separated. These distinct emission sources are seen in figure 3(c).

The objective performs both roles of focusing the pump to a spot and collecting the PL through a window in the cryostat. The infinity corrected, long working distance objective has 10× magnification and a numerical aperture of 0.42. The sample was fixed, and the objective was mounted on XYZ translation stages. Three mirrors were mounted to the stages in a periscope configuration that maintains a consistent relationship between beams on the optical table and the objective, despite the movement of the objective. A visible camera was used to navigate the sample and position the pump spot. More detail on visible microscopy, motion control, spot size control, and IR imaging can be found in the Supplementary Information section.

The procedure for discerning waveguide-coupled PL with this setup is as follows. Initially, the setup was aligned by collecting the free-space-coupled PL from an unpatterned, implanted region. The pump was positioned over an implanted MRR and fine adjusted such as to maximize the free-space-coupled emission signal from the MRR. Then, a razor blade in an image plane is moved to block the center of the field of view, seen in figure 3(d). The blade is translated by a further 500 μm corresponding to 50 μm on chip. Figure 3(b) shows the beams when in the blocking configuration, and figure 3(e) shows the corresponding IR image. Once the center is blocked, mirrors before the spectrometer are adjusted to optimize the offset signal. In this blocking configuration, only the offset PL from the grating coupler is observed.

The optical cryostat is based on a modified cryopump. The top flange of the cryopump chamber was removed, and a machined brass post and brass sample holder were attached to the 15 K stage of the cold

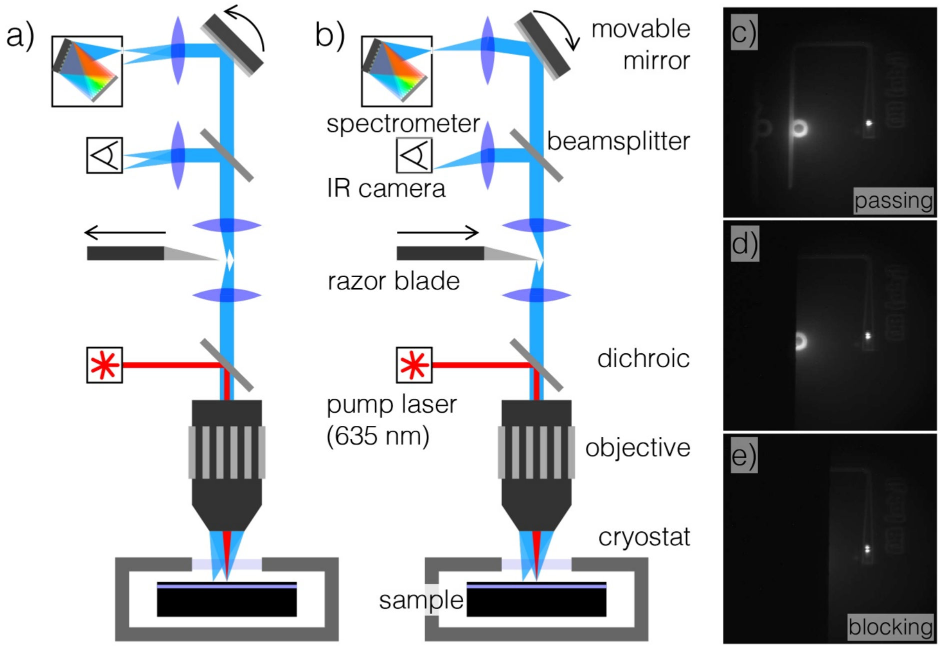


Figure 3. Simplified experimental diagram and visualization of the center PL blocking technique. The configuration can be non-blocking (a) or blocking (b). A pump laser (red) is focused onto a W-center device. Infrared (IR) PL is collected by the same objective. The returning PL transmits through a dichroic mirror and a razor blade assembly. Then the PL is split to an IR camera and spectrometer. When the central PL is blocked, the mirrors preceding the spectrometer are adjusted such that the offset component enters the slit of the spectrometer. (c-e) Verification of the razor blade technique for isolating offset PL. (c) Razor non-blocking corresponding to (a); (d) razor halfway across MRR; (e) razor fully blocking, corresponding to (b). Data is collected in the blocking state.

head. A radiation shield canister with a window was attached to the 80 K stage. Lastly, a new top flange with a window was put in place to seal the vacuum chamber. The cold head is driven by a closed-cycle helium compressor. A silicon diode thermometer is attached to the bottom of the sample holder and read a minimum temperature of 19.8 K. The cryopump motor causes substantial vibration, meaning that optical measurements of small devices cannot be taken while it is running. We estimated a sample vibration displacement between 50 and 100 μm. We addressed this vibration problem by turning the compressor off during fine alignment and measurement. While it is off, the temperature begins to increase, reaching 40 K in approximately 12 minutes. W-center brightness drops by 12% between 28 K and 40 K and then drops off precipitously above 45 K [13, 17]. All measurements in this work were taken between 28 K and 40 K.

4. Results

Waveguide-coupled spectra from MRRs with six different radii are shown in figure 4. These spectra (red) are compared to that of the control region (blue), which was an unpatterned, implanted SOI region. Emission at the zero phonon line (ZPL) at 1218 nm is 21-times brighter than the peak of the sideband; however, the sideband emission integrated over 1225 nm–1275 nm accounted for 55% of the total emission. In this measurement, the sideband was 20 dB above background noise, allowing us to spectrally resolve emission and transmission spectra over a broad, 50 nm wavelength range. It therefore can be used as a diagnostic technique for a wide range of cryogenic silicon photonic devices.

The collected PL is brighter at wavelengths that resonate with the microresonator and nearly undetectable off-resonance. This result indicates that only the PL emitted into the resonant modes couples to the bus waveguide. Preferential waveguide coupling does not necessarily mean preferential emission into the resonant modes, as in the sense of stimulated emission or Purcell enhancement, of which we did not observe conclusive evidence. The shape of the envelope of the peaks closely follows that of the W-center sideband from the control region. This indicates that the wavelength dependence of evanescent couplers and GC are relatively flat over the window. The data are not normalized, and it is coincidental that the amplitudes (dependent on multiple unmeasured factors, discussed below) also match. In figure 4(f), the peak

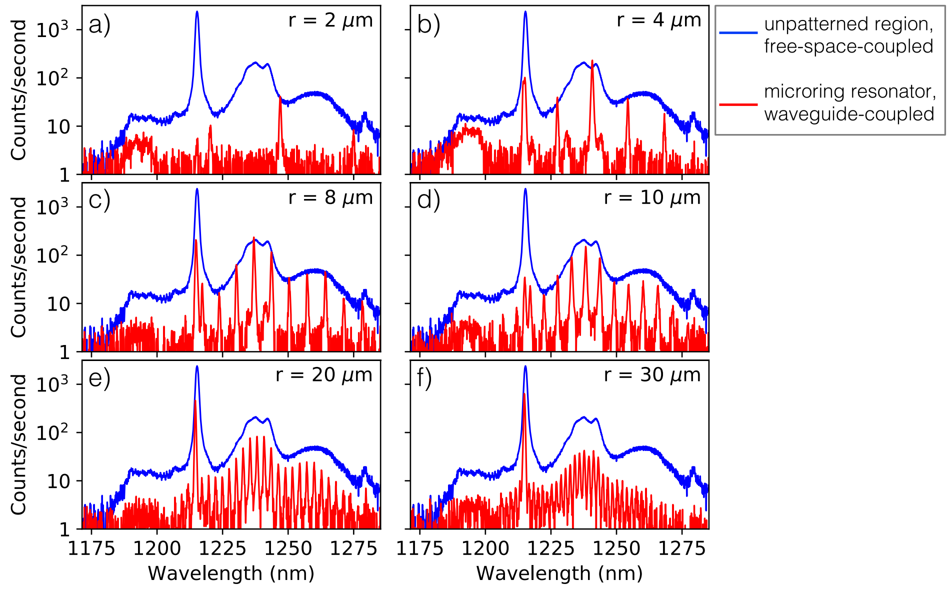


Figure 4. Comparison of free-space-coupled PL from an unpatterned control region (blue curves) to waveguide-coupled PL from the GCs connected to MRRs of different radii (red curves). Multiple peaks at evenly spaced frequencies are indicative of multiple MRR resonances. As MRR radius increases, the free spectral range decreases. The zero phonon line amplitude depends on where its wavelength falls relative to a resonator mode.

amplitudes are significantly weaker because the MRR diameter was much larger than the focused pump spot. The pump was defocused so that more of it reaches the MRR waveguide; however, most of the pump power still passed through the unimplanted center of the largest MRRs.

The zero phonon line was visible in every sample, but its amplitude varied considerably. Its amplitude depends on how close a MRR resonance falls to 1218 nm. As a result of resonance sensitivity to fabrication variation, it is difficult to control the absolute resonant wavelengths [25]. On the other hand, the relative spacing between resonances, i.e. free spectral range (FSR), is precisely determined by MRR radius according to the formula

$$\text{FSR} = \frac{\lambda^2}{2\pi r \cdot n_g}, \tag{1}$$

where n_g is the group index, and λ is wavelength. The best fit of the measured FSRs from figure 4 in the 1225 nm–1250 nm range yields $n_g = 4.51$ with a fit residual of $R^2 = 0.991$, meaning that there is close agreement between this model and the data. Previous measurements of similar MRRs, taken using a transmission spectrum analyzer and fiber alignment setup, yielded a group index of 4.33 ± 0.03; however, that measurement was at room temperature and without W centers.

Figure 5(a) shows a high resolution waveguide-coupled spectrum of the $r = 8$ μm MRR. A spectrometer grating with 1,200 grooves mm^{-1} was used for this measurement compared to the spectrometer grating used for figure 4 with 300 grooves mm^{-1}. The full-width half-maximum (FWHM) of the center peak was measured to be $\Delta\lambda = 0.23$ nm, which is near the resolution limit of the detector array. The quality factor of this feature is 5,300. Since any kind of misalignment or defocus can affect the actual resolution limit, it is possible that this Q factor is underestimated. The intrinsic Q of this MRR was at most 55,000 due to waveguide propagation loss. The extrinsic Q is determined by the combined round trip propagation loss plus ring-to-waveguide coupling coefficient, which was simulated to be $\kappa^2 = 0.121$. The extrinsic Q is therefore expected to be at most 7,700. Measurement and simulation place relatively tight bounds on the actual extrinsic Q between 5,300–7,700. In [24], measurements of a microdisk on the same setup and sample found resonances with Q factors up to 7,160. The higher Q in that work can be explained because higher-radial-order microdisk modes couple more weakly to the bus waveguide and interact less with any etched sidewalls. From the Q factors, we can estimate the coupling efficiency from the optical modes to the grating coupler. Resonator coupling efficiency is $1 - Q_e/Q_i$, where Q_e is extrinsic (5,300: measured directly), and Q_i is intrinsic (55k: estimated from loss measurements). Half of the viable PL propagates

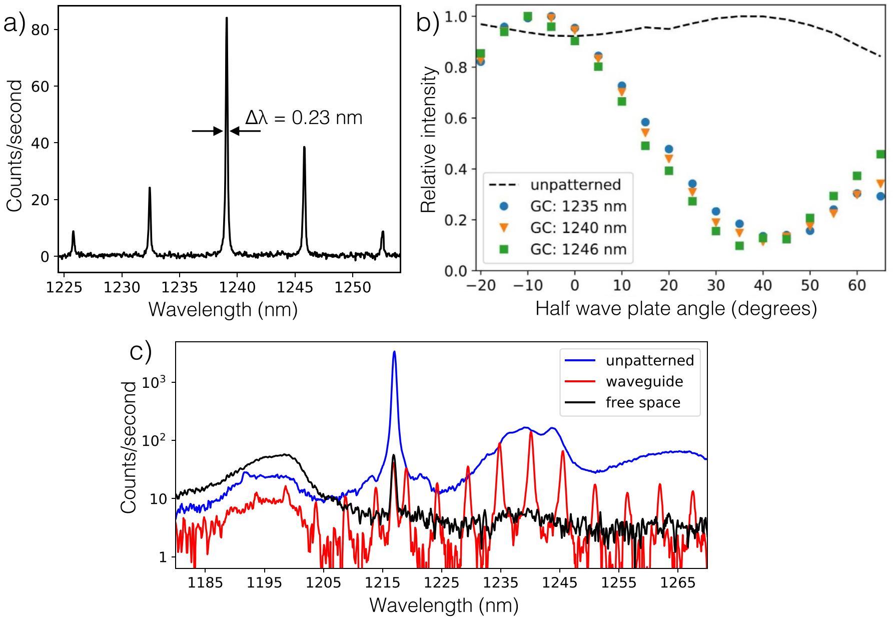


Figure 5. (a) High resolution spectrum of the sideband of the MRR with $r = 8\mu$m (figure 4(c)). The Q-factor of the central peak is at least 5,300. (b) Polarization dependence of light collected from the grating coupler (GC) compared to unpatterned control region. The GC curves are derived from three peaks of the $r = 10$ μm microring (figure 4(d)). Each curve is normalized to its maximum, and the values collected from the GC are first normalized by the corresponding unpatterned value. Note that (a) and (b) are different devices. (c) Comparison of waveguide-coupled and free-space-coupled PL from the $r = 10$ μm microring. The ZPL is visible in the free-space emission, but resonance features are not.

counter-clockwise and travels away from the GC to a waveguide taper. The waveguide from MRR and GC is 200 μm long, resulting in 0.36 dB of loss. The net efficiency between light in a cavity mode and the front of the GC is thus estimated to be 43%.

Figure 5(b) shows the PL polarization. To measure polarization, a half wave plate and polarizer were placed before the razor blade assembly. The allowed polarization angle is twice the half wave plate angle. Values labeled ‘GC’ refer to the amplitudes of different peaks in the waveguide-coupled sideband of the $r = 10$ μm MRR (figure 4(d)). Unpatterned values refer to the ZPL amplitude of an unpatterned control region. The degree of polarization of the GC emission was substantially stronger (88% depth) than that of the control (7.8% depth), which is corroborative because the GC is designed to direct only TE light vertically towards the objective. A half wave plate angle of zero corresponds to S-polarization (E-field perpendicular to the dichroic surface is nonzero). The sample was mounted roughly parallel to the setup such that TE polarized light in its waveguides ended up as S-polarized in the setup. The maximum of GC traces actually occurred at $-10° \pm 5°$. Most likely, this means that the sample was rotated by this amount away from parallel.

Figure 5(c) compares waveguide-coupled and free-space-coupled emission for the $r = 10$ μm MRR. In the sideband, there was no periodic resonance structure in the free-space-coupled component. This confirms that PL was not preferentially emitted into resonator modes, rather, that PL which was emitted into the modes preferentially coupled to the bus waveguide. For this radius and above, most of the pump power missed the MRR waveguide implanted with W centers and instead went through the ring center. As a result, the free-space-coupled ZPL was dimmer than the unpatterned ZPL. This contrasts with similar measurements of a microdisk in [24] wherein the microdisk was implanted throughout, and free-space emission was slightly brighter than the control.

A broad emission band appeared at 1186–1202 nm, which we believe to originate from unimplanted silicon. In this sample, the silicon slab more than 3 μm from waveguides was not etched in order to reduce electron beam write times (see figure 2(b)), which means, as an artifact of fabrication, there was an unetched, unimplanted silicon disk in the center of the MRR that was, for the large MRRs, illuminated with the majority of the pump power. This band was visible in all three traces, but its brightness relative to the ZPL increased by two orders-of-magnitude when comparing the free-space spectra between the MRR and unpatterned control. This feature was strongly non-linear and sensitive to pump power and spot size, explaining why it is prominent here (21 mW pump) but not in most measurements above (11 mW pump). In both control and waveguide-coupled traces, it was about one order-of-magnitude dimmer than the W-center sideband; nevertheless, we can observe that this component was waveguided as evidenced by the resonances here visible at 1194 nm, 1198 nm, and 1202 nm. Luminescence features at these wavelengths have been observed from other silicon samples [10, 26] and likely originate from a diversity of broadened phonon-assisted transitions and/or other defect states [9].

5. Discussion

Previous work on waveguide-coupled W-center emission used electrical pumping and on-chip detection [14] whereas this work used free-space optical pumping and collection techniques. The integrated and free-space approaches present complementary advantages for research. As a consequence of optical pumping, a whole array of devices can be probed in a single cooldown. Luminescence properties can be studied independent of electrical properties such as ohmic heating. Finally, fabrication is two-step (Si rib etch, Si ion implant) instead of seven-step needed for electrical injection (Si rib etch, Si partial etch, Si ion implant, heavy/light boron, heavy/light phosphorus) [14]. In addition to these optical pumping advantages, optical collection provides other complementary advantages. Superconducting detectors are not required, meaning that operating temperature can be 40 K instead of 1 K. Free-space collection allows use of commercial detection instruments (e.g. spectrometers, high-speed diodes, cameras) without a need for fibers in the cryostat. The presented techniques could thus play a central role in the characterization of cryogenic silicon photonic platforms. As an example, we have observed the FSRs of various MRRs and precisely determined their group velocity by examining the sideband spectrum, while any sort of on-chip detection approach would require an on-chip spectrometer to provide similar information.

In several cases, it would be desirable to make a resonance align with the bright ZPL. This alignment presents a challenge because both the resonances and the ZPL are narrow, and the resonant wavelengths are sensitive to fabrication variability. One approach to resonance alignment are cavities designed with FSRs smaller than the ZPL linewidth, which would ensure that at least one resonance would fall within the ZPL. In this work, the $r = 30$ μm MRR has an FSR that meets the condition; however, this device is much larger than the pump spot size, so nearly all of the pump power passes through the center of the ring, missing the implanted waveguide. We fabricated a $r = 40$ μm device, but it was not measurable for this reason. Next steps could include modifying the ring shape or increasing pump power further. In general, the approach of using small-FSR cavities has the issue that they must be physically large, meaning that they have undesirably large footprint and capacitance. Additionally, without a means of tuning, parasitic index changes (e.g. from electrical pumping) cannot be countered.

Refractive index tuning is a more favorable approach for static and dynamic resonance alignment. Tuning is typically implemented with heaters at room temperature. Below 4 K, however, thermooptic tuning is likely not viable because $\mathrm{d}n/dT$ approaches zero [27]. There are other candidate approaches to refractive index tuning at low temperature, including carrier depletion modulation [28], electromechanical tuning [29], and localizing heating to raise the device temperature to between 4 K and 40 K. Static adjustment of MRR resonances could be realized with defect-mediated or other postfabrication trimming [30–32], although these techniques are not capable of dynamic adjustment.

Investigations into a W-center laser would be of particular interest for further work. We did not observe conclusive evidence of lasing in this study primarily due to the resonance alignment challenge. A silicon laser would have a variety of applications for on-chip optical measurement and non-linear optics, for example, photon pair generation through degenerate four wave mixing [33, 34]. Stimulated emission has been observed from G-center defects in nanopatterned silicon [35]. It remains to be shown whether population inversion—let alone amplification—can be achieved in a waveguide that is implanted with W centers.

Substantial research has been dedicated to study of PL in photonic crystal cavities, which differ significantly from MRR cavities [36–38]. Photonic crystal cavity designs often have a goal of minimizing mode volume, while MRRs make use of their small FSRs to more easily align a cavity resonance to the emitter wavelength. Their FSRs are inversely proportional to their mode volumes, which are the product of circumference and modal cross-section. MRRs are essentially bent waveguides that can be oxide-clad. As compared to air-clad photonic crystals, MRRs integrate more readily with traditional silicon photonic platforms, metal layers for index tuning, and other waveguide-based photonic integrated circuits.

The data in this manuscript might be improved by characterizing the GC response spectrum, but likely only slightly. As seen in figure 2(c) and (d), emission appeared to originate entirely from the front of the GC, and sometimes the emission pattern was bifurcated into two lobes, which indicates that this GC is not optimal. Despite the GC being non-optimal, its response is flat over the phonon sideband of interest, which can be inferred by comparing the shape of the control spectrum to the envelope of GC peaks in figure 4. The current setup would be ideal for GC characterization and optimization in further work: using an array of identical sources coupled to varying GCs, the GC that is optimal for objective-coupled normal emission at 1218 nm at 20 K could be found in a single cooldown.

This work presents several other directions for further study. Optical cavities provide a degree of control over emitter lifetime, as in reference [20]. Source lifetime is a critical parameter for superconducting optoelectronic neural networks [39]. Other silicon defect centers can be induced by lithium, carbon, or copper implants [9]. We find that, in practice, W centers produced by silicon ion implant have been more reliable in terms of brightness uniformity. These other centers possess complementary advantages, and further work could apply the current techniques to their investigation. Another prospective direction would be using the broad sideband for wavelength-division multiplexed (WDM) sources. Although the ZPL peak is 21-times brighter than the peak of the sideband, the sideband contains 55% of total emission power when integrated from 1230 nm to 1275 nm. This power may prove sufficient to implement multiwavelength neuromorphic silicon photonic architectures [40] with monolithically integrated light sources. Finally, engineering W-center density to be on the order of λ^{-3} (one emitter per wavelength volume) could be a route to silicon single-photon sources; however, at this time, the density of W centers has not been measured, let alone as a function of implant parameters.

In conclusion, we have used micro-PL measurements with a spatial-filtering technique to unambiguously measure microcavity- and waveguide-coupled luminescence from silicon light sources. These all-silicon sources were locally created with lithographic patterning and integrated with photonic circuits. We found that light emitted into resonant modes is preferentially coupled to an adjacent waveguide. We have illustrated how this setup can employ the phonon sideband as a tool to characterize cryogenic silicon photonic circuits, extracting device values for FSR, group velocity, quality factor, and polarization. These results and methods open numerous directions for further research of fundamental defect center properties, all-silicon active devices, and large-scale photonic information processing.

We thank R. Boutelle and K. Silverman for helpful discussions on the experimental setup. We thank S.-P. Yu and N. Sanford for editorial input. This is a contribution of NIST, an agency of the U.S. Government, not subject to copyright.

ORCID iDs

A N Tait https://orcid.org/0000-0002-9774-4131
S M Buckley https://orcid.org/0000-0003-2809-9287
A N McCaughan https://orcid.org/0000-0002-8553-6474
S Papa Rao https://orcid.org/0000-0002-5893-2535
S W Nam https://orcid.org/0000-0002-4472-4655
J M Shainline https://orcid.org/0000-0002-6102-5880

References

[1] Pérez D, Gasulla I and Capmany J 2018 *Opt. Express* **26** 27265–78
[2] Shainline J M, Buckley S M, Mirin R P and Nam S W 2017 *Phys. Rev. Applied* **7** 034013
[3] Tait A N, de Lima T F, Zhou E, Wu A X, Nahmias M A, Shastri B J and Prucnal P R 2017 *Sci. Rep.* **7** 7430
[4] Liang D and Bowers J E 2010 *Nat. Photon.* **4** 511–17
[5] Zhou Z, Yin B and Michel J 2015 *Light: Sci. Appl.* **4** e358–e358
[6] Roelkens G, Liu L, Liang D, Jones R, Fang A, Koch B and Bowers J 2010 *Laser Photon. Rev.* **4** 751–79
[7] Liao M, Chen S, Park J S, Seeds A and Liu H 2018 *Semicond. Sci. Technol.* **33** 123002
[8] Bao S *et al* 2017 *Nat. Commun.* **8** 1845
[9] Davies G 1989 *Phys. Rep.* **176** 83–188
[10] Shainline J and Xu J 2007 *Laser Photon. Rev.* **1** 334–48
[11] Brádler K, Dallaire-Demers P L, Rebentrost P, Su D and Weedbrook C 2018 *Phys. Rev.* A **98** 032310
[12] Tan S H and Rohde P P 2019 *Rev. Phys.* **4** 100030
[13] Davies G, Lightowlers E C and Ciechanowska Z E 1987 *J. Phys. C: Solid State Phys.* **20** 191–205
[14] Buckley S, Chiles J, McCaughan A N, Moody G, Silverman K L, Stevens M J, Mirin R P, Nam S W and Shainline J M 2017 *Appl. Phys. Lett.* **111** 141101
[15] Bao J, Charnvanichborikarn S, Yang Y, Tabbal M, Shin B, Wong-Leung J, Williams J S, Aziz M J and Capasso F 2008 Point defect engineered Si sub-bandgap light-emitting diodes *Device and Process Technologies for Microelectronics, MEMS, Photonics, and Nanotechnology IV* vol 6800 ed Tan H H *et al* (Bellingham, WA: SPIE) pp 164–71
[16] Chong W, Yu Y, Yang R D, Liang L, Fei X and Ji-Ming B 2011 *Chinese Phys.* B **20** 026802
[17] Buckley S M, Tait A N, Moody G, Olson S, Herman J, Silverman K L, Rao S P, Nam S W, Mirin R P and Shainline J M 2019 arXiv: 1911.01317
[18] Radulaski M *et al* 2015 *ACS Photonics* **2** 14–9
[19] Shakoor A *et al* 2013 *Laser Photon. Rev.* **7** 114–21
[20] Sumikura H, Kuramochi E, Taniyama H and Notomi M 2014 *Sci. Rep.* **4** 5040

[21] Koseki S, Zhang B, De Greve K and Yamamoto Y 2009 *Appl. Phys. Lett.* **94** 051110

[22] Englund D, Ellis B, Edwards E, Sarmiento T, Harris J S, Miller D A B and Vučković J 2009 *Opt. Express* **17** 15409–19

[23] Coles R J, Prtljaga N, Royall B, Luxmoore I J, Fox A M and Skolnick M S 2014 *Opt. Express* **22** 2376–85

[24] Tait A N, Buckley S M, McCaughan A N, Chiles J T, Nam S, Mirin R P and Shainline J M 2020 Microresonator-enhanced, waveguide-coupled emission from silicon defect centers for superconducting optoelectronic networks *IEEE Optical Fiber Comm. Conf. (OFC)* (https://doi.org/10.1364/OFC.2020.M2K.6)

[25] Lu Z, Jhoja J, Klein J, Wang X, Liu A, Flueckiger J, Pond J and Chrostowski L 2017 *Opt. Express* **25** 9712–33

[26] Lightowlers E C, Canham L T, Davies G, Thewalt M L W and Watkins S P 1984 *Phys. Rev.* B **29** 4517–23

[27] Komma J, Schwarz C, Hofmann G, Heinert D and Nawrodt R 2012 *Appl. Phys. Lett.* **101** 041905

[28] Gehl M *et al* 2017 *Optica* **4** 374–82

[29] Yan X, Wu J, Watt R C, Nantel M K T, Chrostowski L and Young J F 2018 Mechanically tunable photonic crystal cavity with high quality factor and small mode *Conf. on Lasers and Electro-Optics* (San Jose, CA: Optical Society of America) p JTh2A.3

[30] Ackert J J, Doylend J K, Logan D F, Jessop P E, Vafaei R, Chrostowski L and Knights A P 2011 *Opt. Express* **19** 11969–76

[31] Alipour P, Atabaki A H, Askari M, Adibi A and Eftekhar A A 2015 *Opt. Lett.* **40** 4476–9

[32] Chen B, Yu X, Chen X, Milosevic M M, Thomson D J, Khokhar A Z, Saito S, Muskens O L and Reed G T 2018 *Opt. Express* **26** 24953–63

[33] Gentry C M *et al* 2015 *Optica* **2** 1065–71

[34] Savanier M, Kumar R and Mookherjea S 2016 *Opt. Express* **24** 3313–28

[35] Cloutier S G, Kossyrev P A and Xu J 2005 *Nat. Mater.* **4** 887–91

[36] Akahane Y, Asano T, Song B S and Noda S 2005 *Opt. Express* **13** 1202–14

[37] Kassa-Baghdouche L 2019 *Phys. Scr.* **95** 015502

[38] Kassa-Baghdouche L and Cassan E 2018 *Photon. Nanostruct. - Fundamentals Appl.* **28** 32–6

[39] Shainline J M *et al* 2019 *J. Appl. Phys.* **126** 044902

[40] Tait A N, Nahmias M A, Shastri B J and Prucnal P R 2014 *J. Lightwave Technol.* **32** 3427–39

Tait, Alexander; Buckley, Sonia; Shainline, Jeff; McCaughan, Adam; Chiles, Jeff; Nam, Sae Woo; Mirin, Richard. ”Microresonator enhanced, waveguide coupled emission from silicon defect centers for superconducting optoelectronic networks.” Paper presented at Optical Fiber Conference, San Diego, CA, US. March 08, 2020 - March 12, 2020.

First Direct Measurement of Sub-Nanosecond Polarization Switching in Ferroelectric Hafnium Zirconium Oxide

X. Lyu[1], M. Si[1], P. R. Shrestha[2], K. P. Cheung[2] and P. D. Ye[1,*]
[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA, *email: yep@purdue.edu
[2]National Institute of Standards and Technology, Gaithersburg, USA

Abstract—In this work, we report on an ultrafast direct measurement on the transient ferroelectric polarization switching in hafnium zirconium oxide with a crossbar metal-insulator-metal (MIM) structure. A record low sub-nanosecond characteristic switching time of 925 ps was achieved, supported by the nucleation limited switching model. The impact of electric field, film thickness and device area on the polarization switching speed is systematically studied.

I. Introduction

Due to the fast speed [1-7], high retention/endurance [8, 9] and CMOS compatible process, ferroelectric (FE) hafnium oxide (HfO_2), such as hafnium zirconium oxide (HZO) [10], has been the promising material for various ferroelectric device applications such as ferroelectric memory (FeRAM), ferroelectric field-effect transistors (Fe-FETs) [1,2,4] and negative-capacitance FETs (NC-FETs) [11-13]. In all these devices, ferroelectric polarization switching speed is a critical factor, which directly determines the working speed of the device. The reported ferroelectric polarization switching times in literatures broadly scatter from few ns to µs [1-7]. Recently, it was studied that the RC time constant in the measurement system can be one of the key problems to under-estimate the intrinsic fast switch speed of ferroelectric hafnium oxide, leading to a broad distribution of polarization switching time [7]. By excluding the impact of RC, 10 ns full polarization switching or few ns characteristics switching time (t_0) were achieved [7]. The fast, scalable in dimension and non-volatile properties enable HfO_2 based Fe-FET to be a strong energy-efficient candidate for non-volatile memory applications and even for the faster volatile memories [14]. However, to replace volatile memories like dynamic RAM (DRAM) or even static RAM (SRAM), sub-ns or GHz operational speed is required and has not been demonstrated in FE HfO_2 system directly.

In this work, the transient ferroelectric polarization switching dynamics in HZO is directly measured by an ultrafast pulse measurement setup. A record low characteristic switching time down to 925 ps is achieved for the first time in sub-ns range, extracted by a nucleation limited switching (NLS) model [15]. A W/HZO/W crossbar capacitor structure and fabrication process are developed to scale down the device area to few µm^2. RC time constants are carefully measured by both capacitance and resistance measurements directly. It is demonstrated that the capacitor area has a significant impact on polarization switching time not because of the RC time constant but due to the intrinsic area dependence. The intrinsic area dependence is ascribed to the multi-grain polycrystalline nature of the FE HZO thin film. The electric field dependence and film thickness dependence are also systematically studied.

II. Experiments

A photo image of the fabricated W/HZO/W crossbar MIM structure is shown in Fig. 1. W is used for both top and bottom electrodes. Ti/Au pad is used on top of bottom W electrode to prevent HZO thin film formation during the atomic layer deposition (ALD) process and also reduce the probe contact resistance. An insulating sapphire substrate is employed to reduce the parasitic effects and signal reflection. The device fabrication process is described in Fig. 2(a). The starting point is an insulating sapphire substrate. W sputtering followed by CF_4/Ar dry etch was used to pattern the bottom W electrodes. Ti/Au contact pads were then defined by photo lithography, e-beam evaporation and a lift-off process. HZO (Hf:Zr=1:1) was deposited by ALD at 200 °C using TDMAHf ($[(CH_3)_2N]_4Hf$), TDMAZr ($[(CH_3)_2N]_4Zr$) and H_2O as Hf, Zr and O precursors. The top W electrodes were defined by sputtering and a lift-off process. A 500 °C rapid thermal annealing was then performed in N_2 for 60 s. Fig. 2(b) and 2(c) shows the cross-sectional schematic of fabricated capacitor and the TEM image on FE HZO. Fig. 3 shows the polarization versus electric field (P-E) hysteresis loop of a crossbar MIM capacitor with 8 nm HZO, showing well-behaved ferroelectric properties. Fig. 4 shows the capacitance-voltage (C-V) characteristics of a crossbar MIM capacitor with 15 nm HZO, with a typical butterfly-shape ferroelectric hysteresis. Fig. 5 is a circuit diagram of the ultrafast pulse measurement setup. A high speed pulse generator along with high speed and high power amplifier (leading to maximum of ~9 V and 300 ps rise time in this work) and a 80 GS/s oscilloscope were used to generate the voltage pulses and monitor the transient switching current. Impedance matched probes and pick-off tee were used to minimize the signal reflections. A positive-up-negative-down (PUND) pulse sequence was applied to directly measure the polarization switching current [7, 16], as shown in Fig. 6, highlighting the preset pulse, switching pulse and non-switching pulse, as also pointed out in Fig. 3. Fig. 7 shows the resistances and the corresponding RC time constant versus different areas, among which the smallest RC constant is below 200 ps. The RC time constants are calculated using experimentally measured resistances and capacitances.

III. Results and Discussion

The voltage pulses with 300 ps rise-time and the corresponding transient current responses are shown in Fig. 8 for both switching (pulse 1) and non-switching (pulse 2) pulses in the PUND measurement. The current response to switching pulse

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Cheung, Kin (Charles); Ye, Pei. "First Direct Measurement of Sub-Nanosecond Polarization Switching in Ferroelectric Hafnium Zirconium Oxide." Paper presented at IEEE International Electron Devices Meeting (IEDM), SAn Francisco, CA, US. December 07, 2019 - December 11, 2019.

(I_{pulse1}) is wider compared to the current response to non-switching pulse (I_{pulse2}) due to the extra polarization switching current. Fig. 9 plots the transient current I_{pulse1}, I_{pulse2} and I_{FE} in the same time scale. The net polarization switching current I_{FE} is calculated by $I_{FE}=I_{pulse1}-I_{pulse2}$. The transient polarization charge density is calculated by the integration of I_{FE} ($P = \int I_{FE}\, dt$), as shown in Fig. 10. A nucleation limited switching model was applied for the multi-grain polycrystalline HZO to extract the characteristic switching time constant. In NLS model for a single grain element, the polarization switching dynamics follow the equation as $P = P_S(1-\exp(-\left(\frac{t}{t_0}\right)^{\beta}))$, where β is an exponential parameter and t_0 is a voltage-dependent characteristic switching time [5,15]. β=2 is used in this work due to 2-dimentional nature of thin film. Sub-ns polarization switching was achieved in 15 nm HZO at 9 V and with a device area of 8.4 μm^2, where the extracted characteristic switching time is 925 ps, as shown in Fig. 10.

The transient FE switching currents for 15 nm HZO capacitors with different areas (Voltage=9 V, E=0.6 V/nm) are plotted in Fig. 11 and the corresponding transient polarization charge densities are shown in Fig. 12. There is a clear area-dependent switching speed as indicated in Fig. 13. The RC time constants for these devices (as shown in Fig. 7) are about 5 times smaller than the characteristic switching times, suggesting the measured switching speed is not limited by the RC effect. Therefore, the ferroelectric polarization switching has an intrinsic area dependence. It is understood that the polarization switching process in polycrystalline FE thin film is a sequential switching process of each elemental grain, according to the nucleation limited switching model, as shown in Fig. 14. At scaled device size, the grain number is significantly reduced so that the polarization switching speed is improved.

Measurements of I_{FE} on 10 nm HZO capacitors under different electric fields are shown in Fig. 15, showing a clear electric field dependence. The corresponding transient polarization charge densities are shown in Fig. 16. Fig. 17 shows the extracted characteristic switching time versus electric field for 10 nm HZO, showing faster polarization switching under higher electric field. Fig. 18 and Fig. 19 show the thickness-dependent transient polarization switching current and transient polarization charge density. The HZO capacitors of each thickness are measured at maximum voltage allowed. The corresponding extracted characteristic time constants versus HZO thickness is shown in Fig. 20, from which the trend of thickness dependence is clear that thicker films are faster. Fig. 21 is a summary of the thickness-dependent remnant polarization (P_r).

The ferroelectric polarization switching time of the FE HZO crossbar capacitor is benchmarked with the reported fastest polarization switching times on FE HfO_2 in literatures, as shown in Table. I. The experimental devices have different structures including capacitor structure (MIM) and FeFET structure, different FE HfO_2, different electrode materials including W, WN, TiN and Ni, different thicknesses from 8 nm to 15 nm and different applied voltage so that under different electric fields. Fig. 22 shows the switching time versus electric field from the benchmark Table I. In this work, a sub-ns (925 ps) record low full polarization switching time is achieved for the first time.

IV. CONCLUSION

In summary, we report on a record low sub-ns characteristic switching time of 925 ps on FE HZO, by a direct ultrafast measurement of transient polarization switching current. The impact of electric field, film thickness and device area on polarization switching speed is systematically studied. The ferroelectric switching speed is significantly improved compared to previous reports and more importantly is approaching GHz regime, suggesting FE HZO to be a promising and competitive high-speed non-volatile memory technology and has the potential even to replace the DRAM and SRAM for high-density, high-speed, and both on-chip and off-chip memories.

ACKNOWLEDGMENT

The authors gratefully acknowledge valuable discussions from J. P. Campbell and S. Salahuddin, technical support from X. Sun and H. Wang on TEM imaging, and the funding support from DARPA/SRC JUMP ASCENT Center.

REFERENCES

[1] W. Chung *et al.*, "First Direct Experimental Studies of $Hf_{0.5}Zr_{0.5}O_2$ Ferroelectric Polarization Switching Down to 100-picosecond in Sub-60mV/dec Germanium Ferroelectric Nanowire FETs," *IEEE Symposium on VLSI*, pp. T89-T90, 2018.

[2] E. Yurchuk *et al.*, "Impact of Scaling on the Performance of HfO_2-Based Ferroelectric Field Effect Transistors," *IEEE Trans. Electron Devices*, vol. 61, pp. 3699-3706, 2014.

[3] H. K. Yoo *et al.*, "Engineering of Ferroelectric Switching Speed in Si Doped HfO_2 for High-Speed 1T-FERAM Application," *IEEE International Electron Devices Meeting*, pp. 481-484, 2017.

[4] S. Dunkel *et al.*, "A FeFET based super-low-power ultra-fast embedded NVM technology for 22nm FDSOI and beyond," *IEEE International Electron Devices Meeting*, pp. 485-488, 2017.

[5] C. Alessandri *et al.*, "Switching Dynamics of Ferroelectric Zr-Doped HfO_2," *IEEE Electron Device Lett*, vol. 39, pp. 1780-1783, 2018.

[6] X. Lyu *et al.*, "Ferroelectric and Anti-Ferroelectric Hafnium Zirconium Oxide: Scaling Limit, Switching Speed and Record High Polarization Density," *IEEE Symposium on VLSI*, pp. T89-T90, 2019.

[7] M. Si et al., "Ultrafast Measurements of Polarization Switching Dynamics on Ferroelectric and Anti-Ferroelectric Hafnium Zirconium Oxide," *Appl. Phys. Lett.* (under review), 2019.

[8] N. Gong *et al.*, "Why is FE–HfO_2 more suitable than PZT or SBT for scaled nonvolatile 1-T memory cell? A retention perspective," *IEEE Electron Device Lett*, vol. 37, pp. 1123-1126, 2016.

[9] K. Ni *et al.*, "Critical role of interlayer in $Hf_{0.5}Zr_{0.5}O_2$ ferroelectric FET nonvolatile memory performance," *IEEE Trans. Electron Devices*, vol. 65, pp 2461-2469, 2018.

[10] J. Muller *et al.*, "Ferroelectricity in Simple Binary ZrO_2 and HfO_2," *Nano Lett.*, vol. 12, pp.4318-4323, 2012.

[11] S. Salahuddin *et al.*, "Use of negative capacitance to provide voltage amplification for low power nanoscale devices," *Nano Lett.*, vol. 8, pp. 405-410, 2008.

[12] M. Si *et al.*, "Steep-slope hysteresis-free negative capacitance MoS_2 transistors," *Nat. Nanotechnol.*, vol. 13, pp. 24-28, 2018.

[13] M. A. Alam *et al.*, "A critical review of recent progress on negative capacitance field-effect transistors", *Appl. Phys. Lett.*, vol. 114, p. 090401, 2019.

[14] K.-H. Kim *et al.*, "Ferroelectric DRAM (FEDRAM) FET with metal/$SrBi_2Ta_2O_9$/SiN/Si gate structure", *IEEE Electron Device Lett.*, vol. 23, pp 82-84, 2002.

[15] J. Y. Jo *et al.*, "Domain switching kinetics in disordered ferroelectric thin films," *Phys. Rev. Lett.*, vol. 99, p. 267602, 2007.

[16] A. Grigoriev *et al.*, "Ultrafast electrical measurements of polarization dynamics in ferroelectric thin-film capacitors," *Rev. Sci. Instrum.*, vol. 80, p. 124704, 2011.

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Cheung, Kin (Charles); Ye, Pei. "First Direct Measurement of Sub-Nanosecond Polarization Switching in Ferroelectric Hafnium Zirconium Oxide." Paper presented at IEEE International Electron Devices Meeting (IEDM), SAn Francisco, CA, US. December 07, 2019 - December 11, 2019.

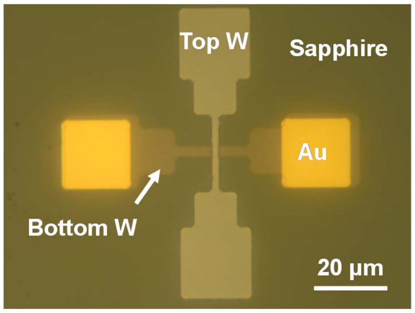


Fig. 1. Photo image of a W/HZO/W crossbar capacitor. An insulator sapphire substrate is used to minimize parasitic capacitance and signal reflection.

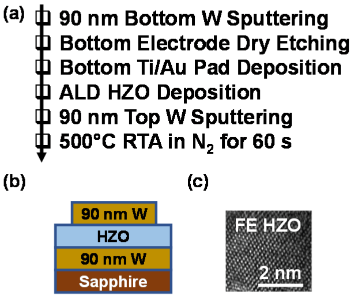


Fig. 2. (a) Device fabrication process flow. (b) Cross-sectional schematic diagram and (c) TEM image of FE HZO.

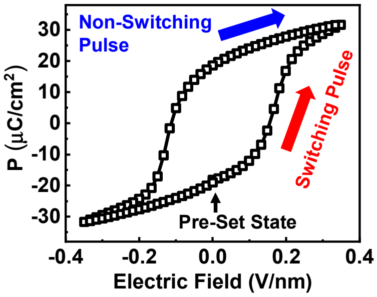


Fig. 3. P-E characteristics of a representative capacitor with 8 nm HZO.

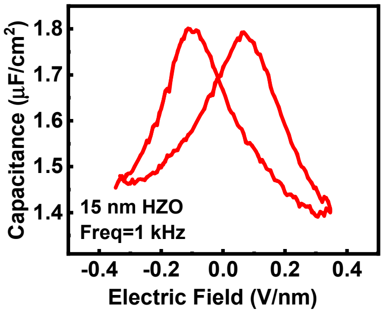


Fig. 4. C-V characteristics of a representative capacitor with 15 nm HZO.

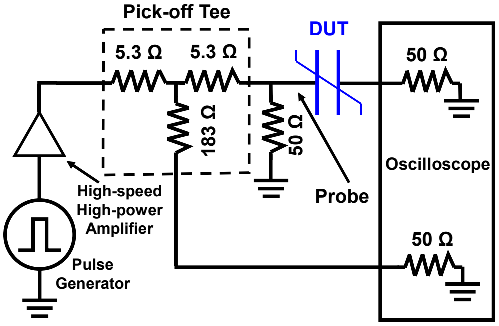


Fig. 5. Circuit diagram of the ultrafast pulse measurement setup with a pulse rise time of ~300 ps.

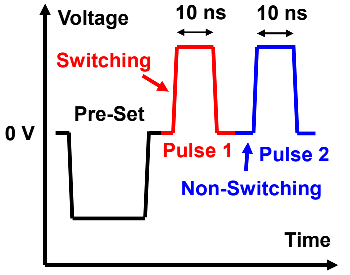


Fig. 6. PUND pulse sequence for transient ferroelectric polarization switching measurement.

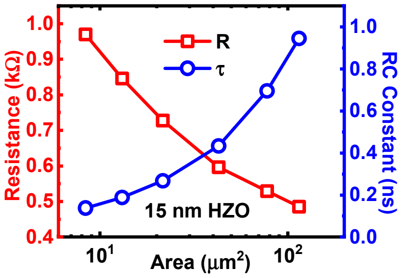


Fig. 7. Resistances and RC time constants of 15 nm HZO MIM devices. RC time constant below 200 ps is achieved and does not affect the speed measurements.

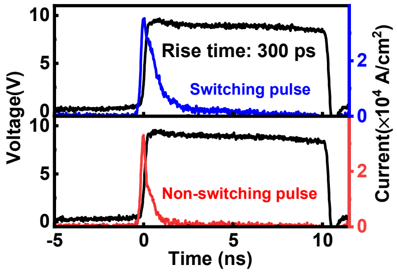


Fig. 8. 10 ns ultrafast pulse input with a 300 ps rise time and the transient current response of the switching (pulse 1) and non-switching (pulse 2) pulses.

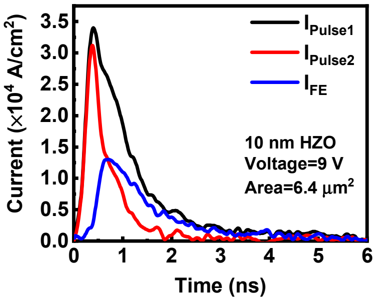


Fig. 9. Transient current I_{pulse1}, I_{pulse2} and I_{FE} of a 10 nm HZO capacitor.

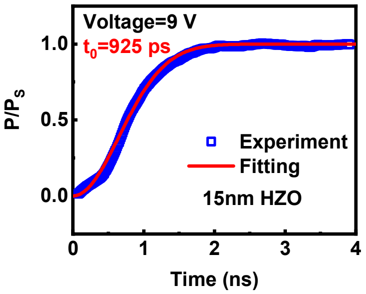


Fig. 10. Normalized transient switched polarization charge density from both experiment and fitting by NLS model. Characteristic switching time is achieved to be 925 ps.

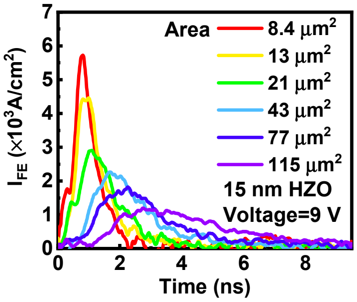


Fig. 11. Transient polarization switching current of 15 nm HZO capacitors with different areas.

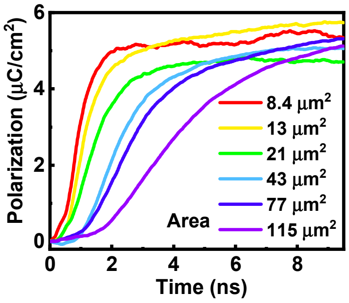


Fig. 12. The corresponding transient polarization charge density by the integration of I_{FE} over time in Fig. 11.

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Cheung, Kin (Charles); Ye, Pei. "First Direct Measurement of Sub-Nanosecond Polarization Switching in Ferroelectric Hafnium Zirconium Oxide." Paper presented at IEEE International Electron Devices Meeting (IEDM), SAn Francisco, CA, US. December 07, 2019 - December 11, 2019.

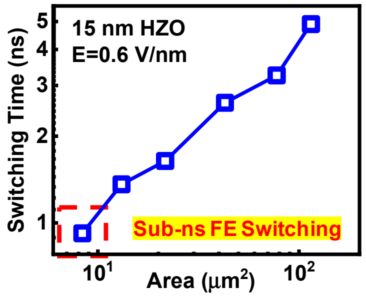


Fig. 13. Area-dependent characteristic switching time in 15 nm HZO.

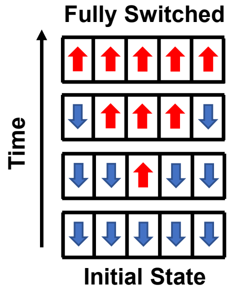


Fig. 14. Illustration of the impact of polycrystalline multi-grains on the FE polarization switching process.

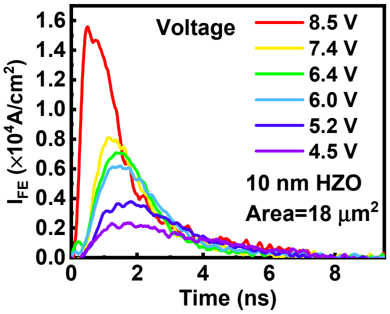


Fig. 15. Transient polarization switching current of 10 nm HZO capacitors under different electric fields.

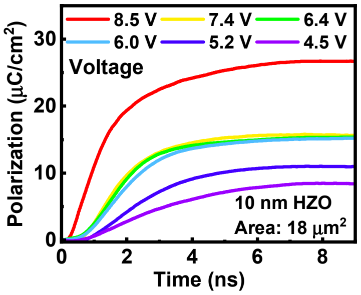


Fig. 16. The corresponding transient polarization charge density from Fig. 15.

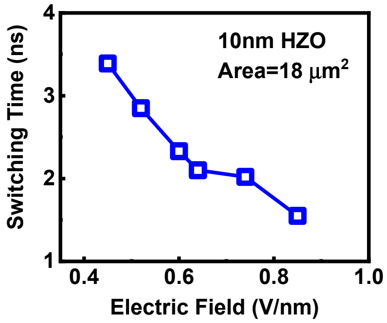


Fig. 17. Electric field dependence of characteristic switching time in 10 nm HZO.

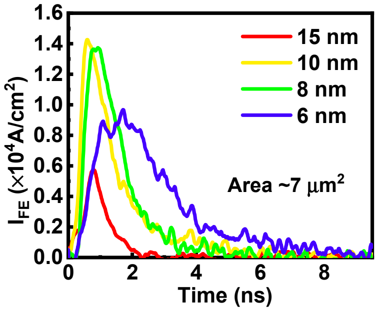


Fig. 18. Transient polarization switching current of HZO capacitors with different thicknesses.

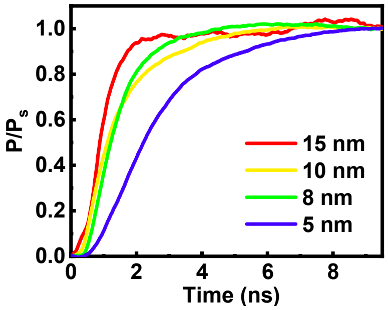


Fig. 19. The corresponding transient polarization charge density from Fig. 18.

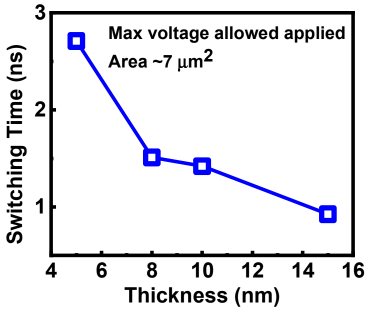


Fig. 20. Fastest characteristic switching time measured in different thicknesses.

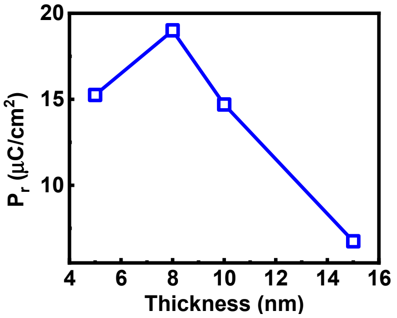


Fig. 21. Thickness-dependent remnant polarization of the HZO capacitor in this work.

	This work Purdue	Ref. 7 Purdue	Ref. 1 Purdue	Ref. 2	Ref. 3	Ref. 4	Ref. 5
Switching Time (ns)	0.925	5.4	3.6	10	100	10	236
FE Oxide Material	HZO	HZO	HZO	Si:HfO$_2$	Si:HfO$_2$	Fe-HK	HZO
Thickness (nm)	15	15	10	9	8	/	8
Structure	MIM	MIM	Ge FeFET	Si FeFET	MIM	Si FeFET	MIM
Electrode Material	W	WN	Ni	TiN	TiN	/	W
Voltage (V)	9	6.7	10	6.5	3	4.2	2.5

Table I. Benchmark of FE HfO2 switching time on MIM and FeFET structures. Sub-ns ferroelectric polarization switching is demonstrated on FE HfO2 for the first time.

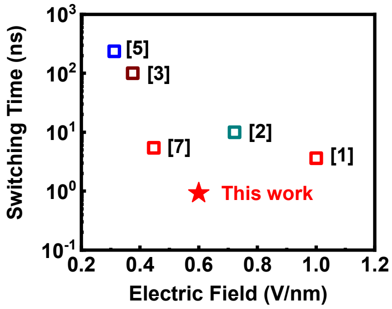


Fig. 22. Benchmark of switching time of FE HfO2 films versus electric field from reported fastest FE switching speed for both MIM and FeFET structures in literatures.

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Cheung, Kin (Charles); Ye, Pei. "First Direct Measurement of Sub-Nanosecond Polarization Switching in Ferroelectric Hafnium Zirconium Oxide." Paper presented at IEEE International Electron Devices Meeting (IEDM), SAn Francisco, CA, US. December 07, 2019 - December 11, 2019.

Proc. of the 11th Model-Based Enterprise Summit (MBE 2020), Gaithersburg, Maryland, USA, March 31 - April 2, 2020

Leveraging standard geospatial representations for industrial augmented reality

Teodor Vernica[1,2], Aaron Hanke[1,3], and William Z. Bernstein[4]

[1] Associate, Systems Integration Division, Engineering Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD, USA
[2] Department of Computer Science, Aarhus University, Aarhus, Denmark
[3] Chair of Engineering Design and CAD, Technische Universitat Dresden, Dresden, Germany
[4] Systems Integration Division, Engineering Laboratory,
National Institute of Standards and Technology, Gaithersburg, MD, USA
{teodor.vernica, aaron.hanke, william.bernstein}@nist.gov

Abstract

Due to its tremendous potential, Augmented Reality (AR) has experienced a recent surge in adoption and integration within the manufacturing enterprise. While industrial AR has been successfully implemented and shown to have significant benefits in a variety of applications, proper use case development, application-specific evaluation, and data interoperability remain open research challenges. In this work, we demonstrate an AR-enabled use case that allows for remote monitoring and inspection of manufacturing systems by overlaying contextual information, such as machine execution data, over the video feed of the manufacturing floor. Additionally, we discuss challenges related to our prototype's implementation and potential opportunities to mitigate such issues through standard indoor geospatial representations.

1 Introduction

In recent years, Augmented Reality (AR) has proven to be a versatile technology that has been leveraged in a multitude of domains including many industrial applications, such as manufacturing planning [5], assembly guidance [14] and maintenance and repair [6] among many others [11]. In this paper, we discuss a new use case for industrial AR that demonstrates remote inspection and monitoring of manufacturing systems by streaming and contextually representing real-time machine process information over the video stream of an Internet Protocol (IP) camera that can be controlled over the network. In doing so, our prototype system provides users, e.g., foremen, operators, and shop managers, with additional capabilities that leverage existing data structures already deployed within smart manufacturing systems. Moreover, by displaying manufacturing information in AR rather than in a strictly digital environment, dynamic elements common to a workshop floor that are much more difficult or even infeasible to track or model are included, such as humans or tools. Our presented prototype also accepts Computer-Aided Design (CAD) models and other virtual 3D objects as inputs that can be layered onto the video feed for a better representation and correlation between the digital (or *as-designed*) models and their physical (or *as-realized*) instances.

During prototyping, we encountered a number of design challenges related to object tracking and camera pose estimation due to the scale of the scene and the rather large distance between the camera and target. These issues are common across other crowded, complex environments akin to busy production floors. This means that traditional tracking methods, such as marker-based tracking, implemented in existing frameworks are not immediately applicable to this scale or are simply infeasible. In other words, popular marker-based recognition methods used in *large target, small field* situations seem to fail when applied to *small target, large field* scenarios.

DOI:https://doi.org/10.6028/NIST.AMS.100-29#page=194

Vernica, Teodor; Hanke, Aaron; Bernstein, William Z. "Leveraging standard geospatial representations for industrial augmented reality." Paper presented at MBE 2020: Model-Based Enterprise Summit 2020, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020.

2 Background

Over the years, numerous different AR tracking and registration techniques have been proposed with various benefits and drawbacks depending on the application context [1, 13].

Fiducial markers have been used in AR applications for decades with diverse designs and implementations [15]. The maturity of marker detection and tracking techniques allows for efficient and reliable pose estimation of the camera that can be computed leveraging the four corners of a marker. In spite of their simplicity and robustness, fiducials have the disadvantage of requiring setup and could cause aesthetic issues, making them inappropriate for certain application environments. In this sense, feature-based markers offer an alternative that trades simplicity for aesthetics, enabled by popular frameworks such as Vuforia[1] and Wikitude[2]. Koch et al. [9] demonstrate how *natural markers* within a building, such as exit signs, can be used for tracking while being seamlessly immersed in the environment.

While both *artificial* and *natural* markers are well-suited for most AR applications, where the markers are relatively close to the camera, success of marker detection and the precision of pose estimation generally decreases with distance. This challenge is evident when applying techniques to wider, more complex environments such as manufacturing floors (often to the point where markers become undetectable). *Artificial* markers seem to perform better than the *natural* markers in such conditions, being more robust to far-field detection and bad camera focus conditions due to their purposeful design.

Even so, most markers were not designed with far-field detection in mind, as most of them report detection ranges of around 3 m for a 20 x 20 cm marker [10]. Cho and Neumann [3] acknowledge this range limitation and present a multiring fiducial design that is able to smoothly zoom-out from near-field to room-sized detection, promising a detection range of up to 4.5 m using a 4 cm diameter circular fiducial. However, at least four of them are required to be in view to calculate the camera's pose. Claus and Fitzgibbon [4] propose a machine learning approach to marker detection, using a marker comprised of four circles on a white background, that shows a significant decrease in error rates compared to square marker detection systems for challenging environment conditions, e.g., bad lighting and far-field detection.

Recently, marker-less methods, such as Simultaneous Localization and Mapping (SLAM) [12], have gained popularity as an alternative to marker-based tracking. Such techniques aim to dynamically build a 3D map of the environment using the already existing natural features without the need of a former setup. This approach is particularly useful for unknown environments or when careful marker placement in the environment is impossible or impractical. Researchers have extended SLAM-based methods to deal with large, complex and dynamic environments. To this extent, Castle et al. [2] present a technique for wide-area tracking by creating multiple distinct maps of different scales that can be used in unison by transitioning from one map to another appropriately. This modular approach has the advantage of only needing to rebuild a subsection of the maps when a change occurs in the target room's configuration. While marker-less tracking techniques can be valuable for tracking the position and orientation of the camera within the environment without the limitation of needing a marker in view at all times, additional work can be required to correctly and automatically register the digital objects in the tracked scene.

Despite its limitations and advances in marker-less tracking methods, marker-based tracking is still effective and commonly used for prototyping purposes due to its robustness and the effortless implementation afforded by frameworks. These limitations can be alleviated in

[1] For more information, visit http://www.ptc.com/en/products/augmented-reality.
[2] For more information, visit https://www.wikitude.com.

combination with marker-less techniques but ultimately there are still challenges when applying these to large, busy environments. These issues were made apparent during the development of our room scale prototypes, which means that they would only be amplified on a larger scale and would require alternative solutions.

3 Camera-Supported Monitoring of Production Systems

To explore opportunities for far-field tracking in the context of industrial AR, we developed a prototype to interface with the National Institute of Standards and Technology (NIST) Smart Manufacturing Systems (SMS) Test Bed[3]. Streaming near-real-time data via a web portal, the SMS Test Bed is representative of a contract manufacturer with a good mix of machine tools. Such an environment offers appropriate testing conditions, e.g., occlusion due to crowded spaces. Our initial testing described here was conducted in the Data Information Visualization and Exploration (DIVE) Lab, recently deployed at NIST.

Our prototype makes use of an off-the-shelf Pan-Tilt-Zoom (PTZ) Internet Protocol (IP) Camera that can stream the video feed and be controlled via the Hypertext Transfer Protocol (HTTP) protocol. A *Unity* desktop application receives the video stream and allows the user to send PTZ commands to the camera, via the keyboard, over the network. Simultaneously, *MTConnect* data generated by computer numerical control (CNC) machines is continuously fetched from the SMS Test Bed, as shown in Figure 1. Quick Response (QR) Codes are used to represent different CNC machines and encode their *MTConnect* Universally Unique IDentifier (UUID). *ZXing*, a barcode processing library, is used for QR Code detection and decoding, while *OpenCV* is used for drawing the detection information on the processed frames. When a QR Code is detected and successfully decoded, the current *MTConnect* data for the respective machine is shown on screen, as shown in Figure 2. In this case, two QR Codes are detected in the frame, representing two different machines: *GFAgie01* and *Mazak01*. Timestamped data corresponding to the two machines is pulled from the SMS Test Bed and displayed on the side-panels next to the video feed as long as they are in view. In doing so, an operator is able to remotely identify which machines are currently producing value or are experiencing downtime.

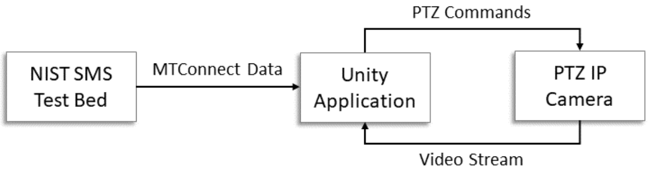


Figure 1: Prototype process diagram.

3.1 Approach Limitations

While our prototype serves as a proof of concept designed to make use of simple ubiquitous technologies, e.g., IP cameras and QR codes, there are some obvious limitations to this approach.

First of all, even more so than AR markers, QR codes are not designed for far-field use, being difficult to detect and especially decode across large distances, unless scaled appropriately, which in itself is often infeasible or impractical. Secondly, there are scalability issues concerning

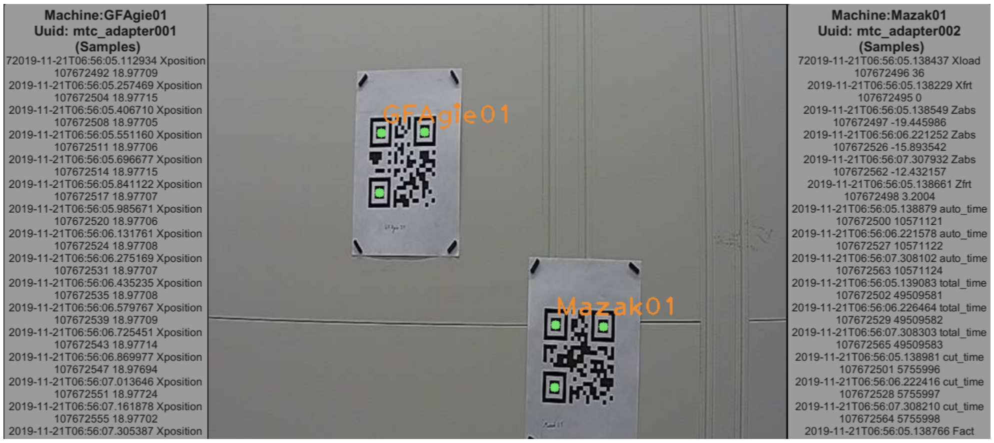
[3] Access to data generated by the SMS Test Bed can be found here: http://smstestbed.nist.gov/.

the number of QR codes in view at any given time, which in turn affects how well users can access data for which they are interested. In other words, the simultaneous detection of multiple machines would also increase the amount of unwanted data on the screen. Additionally, each machine needs to be physically tagged and the markers need to be in the camera's line of sight. Our design allows users to manipulate the line of sight of the camera by accessing its PTZ capabilities. However, detectable markers still (a) need to be oriented orthogonal (or nearly orthogonal) to the camera and (b) cannot be obstructed by other physical objects. This suggests that multiple cameras would be needed to ensure that no machine is obstructed from view. Lastly, while not necessarily a drawback for some use cases, this approach is limited to displaying 2D data over the video feed, given a lack of 3D spacial understanding of the scene.

3.2 Augmenting the Video Feed with Digital 3D Data

Building on the previously described prototype, we present an additional use case showcasing the potential of replacing QR Codes with AR-ready fiducial markers that are more easily tracked by design. Given that the markers can be used to compute the camera pose, 3D objects can be overlaid onto the video stream with an accurate perspective, in addition to the 2D data of the previous use case. This offers the potential of digital models of machines being superimposed over their physical counterparts or displaying any other spatial information in the scene, perhaps in different layers depending on the use context. This is illustrated in Fig. 3, where three CAD models of CNC machines are overlaid on the video feed using the MAXST AR software development kit[4]. A model of the DIVE Lab was created, where three points of interest (the three tables) were mapped. Using this model, the points of interest can be accurately tracked while moving the camera by having a single AR marker in view, thus mitigating some of the issues highlighted by the previous use case such as individual machine tagging.

Note that these prototypes have been implemented and tested in a typical room scale laboratory setting. As described in Section 2, while the prototypes work at this scale, they might not immediately scale appropriately for the desired use case, i.e., a large, crowded, and complex



Machine:GFAgie01
Uuid: mtc_adapter001
(Samples)

```
72019-11-21T06:56:05.112934 Xposition
107672492 18.97709
2019-11-21T06:56:05.257469 Xposition
107672504 18.97715
2019-11-21T06:56:05.406710 Xposition
107672508 18.97705
2019-11-21T06:56:05.551160 Xposition
107672511 18.97706
2019-11-21T06:56:05.696677 Xposition
107672514 18.97715
2019-11-21T06:56:05.841122 Xposition
107672517 18.97707
2019-11-21T06:56:05.985671 Xposition
107672520 18.97706
2019-11-21T06:56:06.131761 Xposition
107672524 18.97708
2019-11-21T06:56:06.275169 Xposition
107672531 18.97707
2019-11-21T06:56:06.435235 Xposition
107672535 18.97708
2019-11-21T06:56:06.579767 Xposition
107672539 18.97709
2019-11-21T06:56:06.725451 Xposition
107672543 18.97714
2019-11-21T06:56:06.869977 Xposition
107672547 18.97694
2019-11-21T06:56:07.013646 Xposition
107672551 18.97724
2019-11-21T06:56:07.161878 Xposition
107672555 18.97702
2019-11-21T06:56:07.305387 Xposition
```

Machine:Mazak01
Uuid: mtc_adapter002
(Samples)

```
72019-11-21T06:56:05.138437 Xload
107672496 36
2019-11-21T06:56:05.138229 Xfrt
107672495 0
2019-11-21T06:56:05.138549 Zabs
107672497 -19.445986
2019-11-21T06:56:06.221252 Zabs
107672526 -15.893542
2019-11-21T06:56:07.307932 Zabs
107672562 -12.432157
2019-11-21T06:56:05.138661 Zfrt
107672498 3.2004
2019-11-21T06:56:05.138879 auto_time
107672500 10571121
2019-11-21T06:56:06.221578 auto_time
107672527 10571122
2019-11-21T06:56:07.308102 auto_time
107672563 10571124
2019-11-21T06:56:05.139083 total_time
107672502 49509581
2019-11-21T06:56:06.226464 total_time
107672529 49509582
2019-11-21T06:56:07.308303 total_time
107672565 49509583
2019-11-21T06:56:05.138981 cut_time
107672501 5755996
2019-11-21T06:56:06.222416 cut_time
107672528 5755997
2019-11-21T06:56:07.308210 cut_time
107672564 5755998
2019-11-21T06:56:05.138766 Fact
```

Figure 2: Real-time MTConnect data is shown for the detected machines.

[4] For more information, visit http://maxst.com.

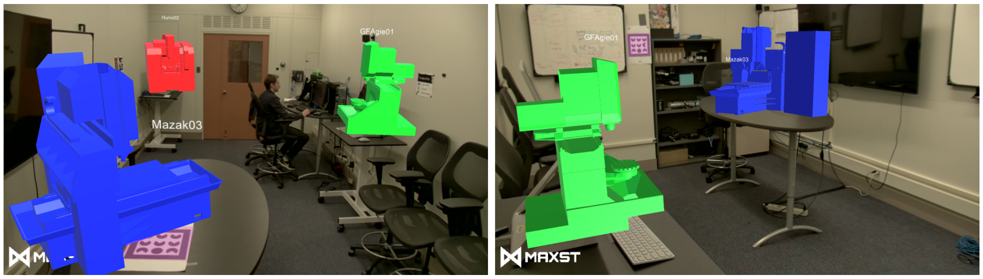


Figure 3: CAD models of CNC machines overlaid on the video feed.

physical environment. Further work and experimentation is required to implement prototype iterations at a larger scale to better understand the full scope of challenges.

4 Future Directions & Opportunities

In this paper, we presented two deployed prototypes that leverage the NIST SMS Test Bed and the DIVE Lab to explore issues related to far-field object tracking in the context of production systems. Based on our preliminary findings, we discuss research directions, including opportunities for standards development.

AR framework developers are pushing towards markerless detection, yet for prototyping and testing purposes, marker-based tracking is still very prevalent throughout industrial implementations. In the case of production systems, where the primary manufacturing assets, such as cranes, industrial robots, and CNC machines, are affixed to a particular location, geospatial definitions can offer more precise data to anchor critical objects in a scene. In other words, rather than relying on techniques such as SLAM to build a feature-map of large area such as a factory floor, we believe that building an as-planned indoor representation might provide additional benefits, such as the ability to include semantics related to the tracked elements and incorporate domain-specific information akin to our MTConnect data streams. Additionally, this approach would alleviate some of the challenges discussed earlier related to far-field marker tracking by removing the need of individual machine tagging and minimizing the number of markers required for the whole scene. This would have the potential for reducing burdens, in terms of both cost, time, and equipment, for testing industrial AR-based prototyping iterations. Furthermore, opportunities exist for the development of measurements methods for the efficiency and appropriateness of as-planned indoor representations.

Moving forward, we plan to explore how richer, pre-defined geospatial representations can influence industrial AR implementation. Based on our early findings, a geospatial representation of a room simplifies the implementation of far-field tracking systems for indoor use. In our prototype, we plan to leverage IndoorGML [8], a standard data format from the Open Geospatial Consortium (OGC)[5]. Specifically designed for formally describing scenarios that require positional data of physical entities inside buildings, IndoorGML[6] provides a framework for geospatial information that relates properties and features of indoor spaces within a flexible framework. We chose IndoorGML for implementation due to existing available tools built around the technology, including an editor for generating IndoorGML documents [7] and

[5] For more information, refer to https://www.opengeospatial.org.
[6] For more information, refer to http://www.indoorgml.net.

an existing link to scene generation in Unity[7], a 3D development platform. Leveraging such geospatial definitions, we hypothesize that (1) less markers would be required for tracking a set of objects, (2) the burden of introducing additional spatially-aware sensors into the predefined environment would be lessened, and (3) such representations coupled with vision-based tracking, like SLAM, would provide more robust object tracking solutions.

Similar to other visualization-driven technologies, industrial AR must overcome a divergence of two traditionally separated standards development communities: (i) the primarily gaming-driven AR frameworks contributed by standards developments organizations (SDOs) such as the Khronos Group and OGC and (ii) data interoperability solutions from SDOs focused on smart manufacturing systems such as the MTConnect Institute[8] and the Open Platform Communications (OPC) Foundation[9]. We believe that our work will provide more guidance and direction for the revision or extension of existing standards and/or opportunities for new standards development. For example, we plan to test how data elements standards by the AR-focused SDOs for affixed and mobile objects (e.g., load-bearing columns and furniture, respectively) relate to analogous manufacturing assets, such as CNC machining centers (affixed) and tooling carts (mobile). We believe that such exploratory tasks will pave the way for the conformance mappings between the two standards communities.

Disclaimer

This work represents an official contribution of NIST and hence is not subject to copyright in the US. Identification of commercial systems in this paper are for demonstration purposes only and does not imply recommendation or endorsement by NIST.

Acknowledgements

We thank Dr. Moneer Helu, Frank Riddick, Dr. Goudong Shao, Dr. Jeremy Marvel, and the anonymous reviewers of the MBE Summit for their valuable feedback to improve this paper.

References

[1] Mark Billinghurst, Adrian Clark, Gun Lee, et al. A survey of augmented reality. *Foundations and Trends® in Human–Computer Interaction*, 8(2-3):73–272, 2015.

[2] Robert O Castle, Georg Klein, and David W Murray. Wide-area augmented reality using camera tracking and mapping in multiple regions. *Computer Vision and Image Understanding*, 115(6):854–867, 2011.

[3] Youngkwan Cho and Ulrich Neumann. Multiring fiducial systems for scalable fiducial-tracking augmented reality. *Presence*, 10(6):599–612, 2001.

[4] David Claus and Andrew W Fitzgibbon. Reliable fiducial detection in natural scenes. In *European Conference on Computer Vision*, pages 469–480. Springer, 2004.

[5] Fabian Doil, W Schreiber, T Alt, and C Patron. Augmented reality for manufacturing planning. In *Proceedings of the Workshop on Virtual Environments*, pages 71–76. ACM, 2003.

[7] For more information, refer to https://unity.com.
[8] For more information, refer to https://www.mtconnect.org.
[9] For more information, refer to https://opcfoundation.org.

[6] Steven Henderson and Steven Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1355–1368, 2010.

[7] Jung-Rae Hwang, Hye-Young Kang, and Jin-won Choi. Development of an editor and a viewer for indoorgml. In *ISA*, pages 37–40, 2012.

[8] Hae-Kyong Kang and Ki-Joune Li. A standard indoor spatial data model—OGC IndoorGML and implementation approaches. *ISPRS International Journal of Geo-Information*, 6(4):116, 2017.

[9] Christian Koch, Matthias Neges, Markus König, and Michael Abramovici. Natural markers for augmented reality-based indoor navigation and facility maintenance. *Automation in Construction*, 48:18–30, 2014.

[10] Pierre Malbezin, Wayne Piekarski, and Bruce H Thomas. Measuring artootkit accuracy in long distance tracking experiments. In *The First IEEE International Workshop Agumented Reality Toolkit,*, pages 2–pp. IEEE, 2002.

[11] SK Ong, ML Yuan, and AYC Nee. Augmented reality applications in manufacturing: a survey. *International Journal of Production Research*, 46(10):2707–2742, 2008.

[12] Gerhard Reitmayr, Tobias Langlotz, Daniel Wagner, Alessandro Mulloni, Gerhard Schall, Dieter Schmalstieg, and Qi Pan. Simultaneous localization and mapping for augmented reality. In *2010 International Symposium on Ubiquitous Virtual Reality*, pages 5–8. IEEE, 2010.

[13] DWF Van Krevelen and Ronald Poelman. A survey of augmented reality technologies, applications and limitations. *International journal of virtual reality*, 9(2):1–20, 2010.

[14] ML Yuan, SK Ong, and AYC Nee. Augmented reality for assembly guidance using a virtual interactive tool. *International Journal of Production Research*, 46(7):1745–1767, 2008.

[15] Xiang Zhang, Stephan Fronz, and Nassir Navab. Visual marker detection and decoding in ar systems: A comparative study. In *Proceedings of the 1st International Symposium on Mixed and Augmented Reality*, page 97. IEEE Computer Society, 2002.

Vernica, Teodor; Hanke, Aaron; Bernstein, William Z. ”Leveraging standard geospatial representations for industrial augmented reality.” Paper presented at MBE 2020: Model-Based Enterprise Summit 2020, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020.

Terahertz Electromagnetically Induced Transparency in Cesium Atoms

Sumit Bhushan*, Oliver Slattery, Xiao Tang, and Lijun Ma†
Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, United States
*Author e-mail address: (*sumit.bhushan@nist.gov, †lijun.ma@nist.gov)*

Abstract: We outline a proposal to realize Electromagnetically Induced Transparency (EIT) with the potential to store Terahertz (THz) optical pulses in Cesium atoms. Such a system, when experimentally realized, has a potential to make Quantum Communication possible with THz signals.

1. Introduction

Terahertz (THz) (0.1 – 10 THz) signals suffer less attenuation and scintillation effects caused by fog, dust, and airborne particles in comparison to Infrared (IR) signals [1, 2]. At the same time, the high level of directionality associated with THz signals reduces the vulnerability to eavesdropping [2]. Additionally, THz signals are compatible with superconducting based quantum computers. Taken together, these factors make THz frequencies an attractive regime for communication purposes and there are several groups conducting research in this field. However, THz quantum communication has not been fully investigated [3, 4]. Here, we present a study in which an ensemble of Cesium atoms is shown to have the potential to store THz signals via Electromagnetically Induced Transparency (EIT). In this scheme, the Cesium atoms are excited to the Rydberg energy levels and EIT occurs within the Rydberg manifold. These energy levels have been chosen because the transition frequencies between the Rydberg levels are of the order of THz and this suggests it is possible to realize EIT based quantum memory for THz signals. We present the theoretical model and simulation results of this scheme.

2. Theoretical Model

The energy level diagram of our model is shown in Fig.1. The ground level (represented as $|g\rangle$) atoms are excited to the Rydberg level $30P_{1/2}$ (represented as $|1\rangle$) by a very strong pulsed Ultraviolet (UV) light at 320 nm and with a repetition rate of GHz order [5]. $|1\rangle$ - $|3\rangle$ and $|2\rangle$ - $|3\rangle$ are the probe and control transitions respectively. The decay from the energy levels $|1\rangle$, $|2\rangle$, and $|3\rangle$ and all other intermediate levels to $|g\rangle$ is compensated by a very strong 320 nm laser which pumps almost all the atoms to $|1\rangle$. Therefore, the four-level system effectively becomes the three-level system shown in Fig.1, which can be used for EIT.

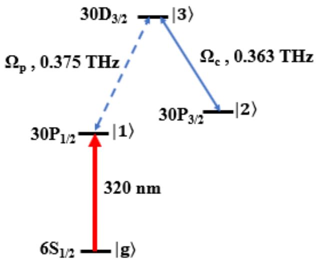


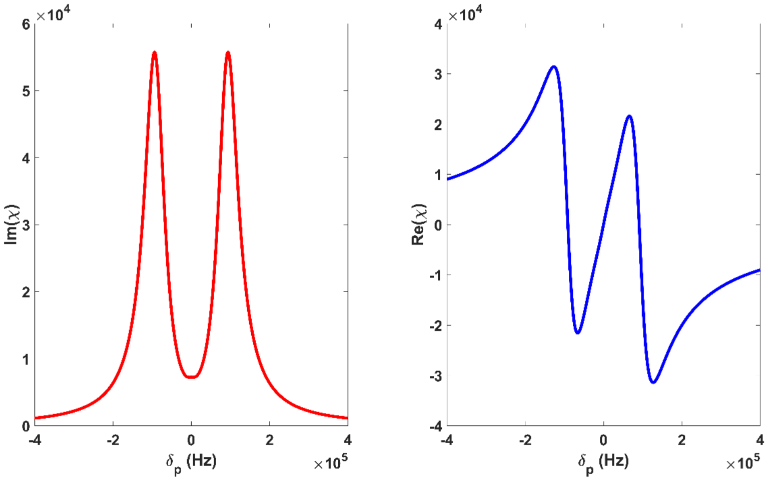
Fig.1 Simplified Cs energy level diagram for our model. Ω_p and Ω_c are the Rabi frequencies of probe and control fields respectively.

The susceptibility of the three-level light-matter system can be written as [6]

$$\chi = i\frac{Nd^2}{\varepsilon_0 \hbar}\frac{1}{\left[\frac{(\Omega_c/2)^2}{\gamma_d - i(\delta_p - \delta_c)} + (\gamma_p - i\delta_p)\right]}$$

Bhushan, Sumit; Slattery, Oliver T.; Tang, Xiao; Ma, Lijun. "Terahertz Electromagnetically Induced Transparency in Cesium Atoms." Paper presented at Frontier in Optics (FIO). September 14, 2020 - September 17, 2020.

where N is the atomic number density, d is the dipole moment associated with probe transition, γ_d is the decoherence rate between $|1\rangle$ and $|2\rangle$, γ_p is the probe decay rate, δ_p (δ_c) is the detuning of probe (control) from its corresponding transition, ε_0 is the permittivity of free space, and ħ is $h/2\pi$ where h is the Planck's constant. The real and imaginary parts of the above equation have been plotted as a function of δ_p as shown in Fig. 2



(a) (b)

Fig.2 (a) The imaginary part of the susceptibility and (b) the real part of susceptibility both plotted as a function of δ_p. The values of the other equation parameters are: $N = 10^{15}$ atoms/m^3, $d = 5.69 \times 10^{-26}$ C.m, $\gamma_p = 0.046$ MHz, $\gamma_d = 0.018$ MHz, $\delta_c = 0$, and $\Omega_c = 4\gamma_p$. The typical EIT features are clearly visible.

The values of various parameters written in the caption of Fig.2 have been either taken from or calculated from the relevant formula given in Ref. [7, 8]. The signal pulse can be spatially compressed inside the atomic ensemble due to the reduction of the probe's group velocity. Note that this reduction in the group velocity is evident from the steep slope around $\delta_p = 0$ in Fig. 2(b). After the signal enters the medium, one can adiabatically turn off the control field and the quantum state of the signals will be mapped on the atomic spin wave. When needed, the signal can be retrieved by turning on the control field. As mentioned above, this system is effectively three levels, hence the readout of the stored pulse will be the same as in a usual three level lambda type EIT based quantum memory. Such a scheme would be useful for quantum memory using THz signals.

3. References

[1] K. Su, L. Moeller, R. B. Barat, and J. F. Federici, "Experimental comparison of terahertz and infrared data signal attenuation in dust clouds," JOSA A **29,** 2360 (2012).
[2] J. Federici and L. Moeller, "Review of terahertz and subterahertz wireless communications," J. Appl. Phys. **107**, 111101 (2010).
[3] J. Wu, B. Lin, J. Wan, L. Liang, Y. Zhang, T. Jia, C. Cao, L. Kang, W. Xu, J. Chen, and P. Wu, "Superconducting terahertz metamaterial mimicking electromagnetically induced transparency," Appl. Phys. Lett. **99**, 97 (2011).
[4] S. Bhushan, V. S. Chauhan, and R. K Easwaran, "Ultracold Rydberg atoms for efficient storage of terahertz frequency signals using electromagnetically induced transparency," Phys. Lett. A **382**, 3500 (2018).
[5] J. E. Sansonetti, "Wavelength, transition probabilities, and energy levels for the spectra of cesium (Cs $_{\mathrm{I}}$ -Cs $_{\mathrm{LV}}$)," J. Phys. Chem. Ref. Data **38**, 761 (2009).
[6] L. V. Hau, S. E. Harris, Z. Dutton, and C. H. Behroozi, "Light speed reduction to 17 metres per second in an ultracold atomic gas," Nature **397**, 594 (1999).
[7] P. Goy, J. M. Raimond, G. Vitrant, and S. Haroche, "Millimeter-wave spectroscopy in cesium Rydberg states. Quantum defects, fine- and hyperfine-structure measurements," Phys. Rev. A **26**, 2733 (1982).
[8] I. I. Beterov, I. I. Ryabtsev, D. B. Tretyakov, and V. M Entin, "Quasiclassical calculations of blackbody-radiation-induced depopulation rates and effective lifetimes of Rydberg nS, nP, and nD alkali-metal atoms with n ≤ 80," Phys. Rev. A **79**, 052504 (2009).

On-Wafer Metrology for a Transmission Line Integrated Terahertz Source

Kassiopeia Smith*, Bryan Bosworth, Nicholas Jungwirth, Jerome Cheron, Nathan Orloff, Christian Long, Dylan Williams, Richard Chamberlin, Franklyn Quinlan, Tara Fortier, and Ari Feldman

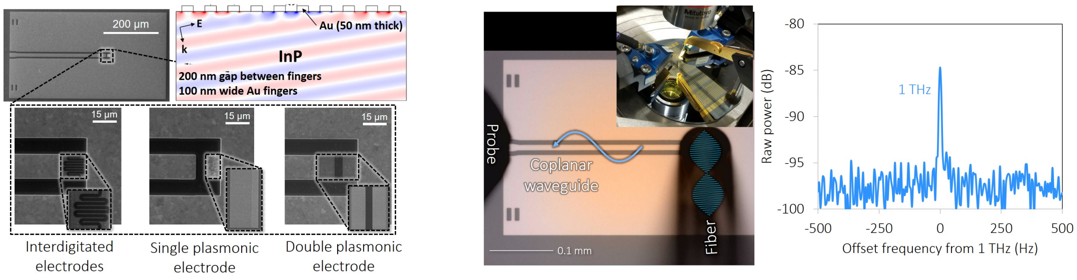
National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305 USA
**send correspondence to: kassiopeia.smith@nist.gov*

Abstract

We developed a measurement system that combines on-wafer metrology and high-frequency network analysis to characterize the response of transmission-line integrated Er-GaAs and InGaAs photomixers up to 1 THz to support the telecommunication and electronics industry.

Radio frequency (RF) test equipment is used in the electronics and telecommunications industry to characterize electronic devices from DC to above 1 THz. Commercial semiconductor foundries need proper high-frequency diagnostic signals to reduce the time-intensive and costly trial-and-error process during development of next-generation wireless communication, computer processing, and other high-speed electronics [1–3]. Complex modulated-signal tests use arbitrary waveforms that are generated by discrete digital steps to construct analog signals, but are fundamentally limited in frequency based on the operating frequency of the digital circuit. Drawbacks to the current RF approach include frequency-banded rectangular waveguides, amplification of phase noise, and poor stability caused by multiplying the RF source to the desired frequency. While interleaving multiple digital sources is one way to overcome the issue, amplitude fluctuation, timing jitter, and noisy signals above 40 GHz remain an issue [4].

Here, we present a broadband optoelectronic source for linear and non-linear network analysis of wafer-based devices and the first steps toward a terahertz synthesizer. We first separate individual frequencies of an optical comb and pass them through a phase-and-amplitude control array that steers the optical frequencies to a fiber. The modulated signals in the fiber are converted into an electrical signal by dividing the frequency down with a photomixers that operate to a terahertz. As a part of this new signal generation paradigm we grew erbium doped GaAs and InGaAs photomixers via molecular beam epitaxy on semi-insulating InP substrates and defined electrodes and co-planar waveguide structures using traditional and electron-beam lithography techniques. We leveraged recent advances in the photomixer field to simulate and incorporate plasmonic electrode structures to enhance terahertz conversion efficiency. Figure 1a shows SEM images of the fabricated devices as well as a multiphysics simulation of the plasmonic enhancement of the optical absorption. Two teeth of an optical frequency comb source excited the photomixers, and a WR1 wafer probe and extender head connected to a vector network analyzer measured the output signal up to 1 THz. Fig. 1b shows the measurement setup and initial data.



(a)

(b)

Fig. 1: a) Optical and SEM micrographs of a transmission line integrated photomixer device and example plasmonic electrode simulation. b) On wafer measurement of devices using frequency comb optical beat

Smith, Kassi; Bosworth, Bryan; Jungwirth, Nick; Cheron, Jerome; Orloff, Nate; Long, Chris; Williams, Dylan; Chamberlin, Richard; Quinlan, Franklyn; Fortier, Tara; Feldman, Ari. "On-Wafer Metrology of a Transmission Line Integrated Terahertz Source." Paper presented at CLEO: Laser Science to Photonic Applications, San Jose, CA, US. May 10, 2020 - May 15, 2020.

There are a large number of industries that rely on electronic devices that operate over a terahertz and currently no way to measure or calibrate devices that operate within that frequency regime. Continuous-wave (CW) terahertz photomixing devices have emerged along side a number of advances to photoconductive materials, electrode design, operating frequency, and conversion efficiency [5]. $In_{0.53}Ga_{0.47}As$ grown on semi-insulating InP substrates allow utilization of low-cost and readily available Er-doped fiber lasers that operate at 1550 nm. While InGaAs is a promising material, there is room to reduce dark current and increase optoelectronic conversion efficiency [6]. InGaAs offers much higher mobilities than the more commonly utilized LT-GaAs, but suffers from lower conversion efficiency at typical bias voltages [7–9]. Implementing advanced growth techniques such as incorporated nanostructures and Bragg reflectors are ways to improve bandwidth, increase signal output, improve resistivity, and add the ability to tune photocarrier lifetimes.

Historically, high-frequency photomixer devices have been intended for terahertz radiation. They are operated by pumping an ultra-fast mode-locked laser onto the electrode contact and measuring the radiated power as a function of bias voltage and optical pump power. Alternatively, fabrication of co-planar transmission-line integrated photomixers with terahertz output represents an innovative advancement for on-wafer metrology. When the photomixer is excited by the heterodyned beat of two tones from an optical frequency comb a high fidelity terahertz signal is generated and can be used as a convenient on-wafer source. Dividing down from the optical regime reduces phase noise at these frequencies, as opposed to traditional multiplication which increases phase noise. Through on-wafer metrology techniques, the propagation constant and dispersion effects of the transmission line can be solved for, which allows the reference plane to be rolled to a plane of interest. For example, when using a photomixer as a source, it will need to be coupled to or embedded within a device or material of interest and the reference plane can be rolled to that interface [10]. Furthermore, when a device is fabricated on an electro-optic substrate, optical network analysis techniques can be used to examine the outgoing and reflected waves of devices under test [11]. The ability to generate pure sine waves up to 1 THz and optically measure outgoing waves and their reflection off a device or material-under-test is a useful measurement tool that will enable development of next-generation technologies with first-pass design success [12].

References

1. S. Aoki, ”Devices, Materials, and Packaging Technologies for Hyperconnected Cloud”, in *Fujitsu Sci. Tech. J.*, 53[2] 3-8 (2017).
2. S. Nellen, B. Globisch, R. B. Kohlhaas, L. Liebermeister, and M. Schell, “Recent progress of continuous-wave terahertz systems for spectroscopy, non-destructive testing, and telecommunication,” in *Terahertz, Rf, Millimeter, and Submillimeter-Wave Technology and Applications*, (2018).
3. M. Malinauskas, A. Zukauskas, S. Hasegawa, Y. Hayasaki, V. Mizeikis, R. Buividas, and S. Juodkazis, ”Ultrafast laser processing of materials: from science to industry,” in *Light-Sci. Appl.*, 5 14 (2016).
4. C. Schmidt, C. Kottke, V. Tanzil, R. Freund, V. Jungnickel, F. Gerfers, ”Digital-to-Analog Converters Using Frequency Interleaving: Mathematical Framework and Experimental Verification,” in *Circuits Syst Signal Proces*, 37 [11] 4929-4954 (2018).
5. S. Preu, G. H. Dohler, S. Malzer, L. J. Wang, and A. C. Gossard, ”Tunable, continuous-wave Terahertz photomixer sources and applications,” in *J. Appl. Phys.*, 109[6] 56 (2011).
6. S. Gupta, J. F. Whitaker, and G. A. Mourou, ”Ultrafast Carrier Dynamics in III-V Semiconductors Grown by Molecular Beam Epitaxy at Very Low Substrate Temperatures,” in *IEEE J. Quantum Electron.*, 28[10] 2464-72 (1992).
7. S. Yang, M. R. Hashemi, C. W. Berry, and M. Jarrahi, ”7.5% Optical-to-Terahertz Conversion Efficiency Offered by Photoconductive Emitters With Three-Dimensional Plasmonic Contact Electrodes,” in *IEEE Transactions on Terahertz Science and Technology*, 4[5] 575-81 (2014).
8. A. Mingardi, W. D. Zhang, E. R. Brown, A. D. Feldman, T. E. Harvey, and R. P. Mirin, ”High power generation of THz from 1550-nm photoconductive emitters,” in *Opt. Express*, 26[11] 14472-78 (2018).
9. N. T. Yardimci and M. Jarrahi, ”Nanostructure-Enhanced Photoconductive Terahertz Emission and Detection,” in *Small*, 14[44] 14 (2018).
10. R. Chamberlin and D. Williams, ”Measurement and Modeling of Heterogeneous Chip-Scale Interconnections,” in *IEEE Trans. Microwave Theory Tech.*, 66 [12] (2018).
11. P. Struszewski, M. Bieler, ”Asynchronous Optical Sampling for Laser-Based Vector Network Analysis on Coplanar Waveguides,” in *IEEE Trans. Instrum. Meas.*, 68 [6] (2019).
12. D. Williams, J. Cheron, R. Chamberlin, T. Dennis, ”Design of an On-Chip mmWave LSNA with Load Pull and Advanced Signal Sources,” in IEEE 2019 93rd ARFTG Microwave Measurement Conference, (2019).

On Data Integrity Attacks against Industrial Internet of Things

Hansong Xu*, Wei Yu*, Xing Liu*, David Griffith†, and Nada Golmie†
*Towson University, USA
Emails: {hxu2,xliu10}@students.towson.edu, wyu@towson.edu
†National Institute of Standards and Technology (NIST), USA
Emails:{david.griffith, nada.golmie}@nist.gov

***Abstract*—Industrial Internet of Things (IIoT) is predicted to drive the fourth industrial revolution through massive interconnection of industrial devices, such as sensors, controllers and actuators, integrating advances in smart machinery and data analytics driven by computing, networking and artificial intelligence techniques. The massive interconnections between industrial devices increase not only information exchange, productivity, and resource efficiency, but also cybersecurity risks and their impact to the IIoT system. In this paper, we investigate the evolutionary process, which captures the natural interactions and evolutions of attackers and defenders, as well as their strategies to launch data integrity attacks and protect the IIoT system, respectively. We leverage the wireless cyber-physical simulator (WCPS) as a realistic IIoT testbed to implement and evaluate strategies at different levels of evolution. The experimental results indicate the effectiveness of defensive strategies in guarding against various forms of data integrity attacks with increase in time and knowledge throughout the evolutionary process.**

***Index Terms*—Industrial Internet of Things, Data Integrity Attacks, Evolutionary Process, and Cyber-Physical System**

I. Introduction

INDUSTRIAL Internet of Things (IIoT), also known as Industry 4.0, is the integration of Internet of Things (IoT) technologies with industrial manufacturing and production. Benefitting from the massive interconnections of IoT, the IIoT is anticipated to improve the productivity, efficiency, and reliability of manufacturing and production processes. From the viewpoint of a cyber-physical system (CPS), the networking, computing, and control subsystems in IIoT are the cyber components, which connect and coordinate physical industrial entities (sensors, controllers, actuators, etc.). Specifically, the networking system enables real-time information transmission to interconnect the physical devices of factories and plants. The computing system brings intelligence to cover all stages of manufacturing and industrial processes through data analytics. The control system controls industrial devices through control loops to ensure reliable system operations.

Despite the great potential for transforming traditional industrial factories to smart factories in Industry 4.0, the integration of IoT will also face cybersecurity risks. The cybersecurity risks are introduced by the massive interconnections among IIoT systems and the attraction of valuable information inherent to industrial critical infrastructures. In addition, the cybersecurity risks could eventually impact the reliable and safe operations of the industrial devices, as well as the manufacturing and production processes. In other words, cybersecurity risks could affect not only the information integrity, but also the safety operations of IIoT systems. To study cybersecurity issues from a CPS perspective, Liu *et al.* [1] investigated the impacts of cyberattacks on cutting-edge smart-world systems, including smart manufacturing, smart grid, and smart transportation. The smart manufacturing system has shown great vulnerabilities to disruption, such as via the Stuxnet worm to gain control of programmable logic controllers (PLCs) [2]. Likewise, the smart grid system has also been shown to be vulnerable to data integrity attacks that could result in disturbance of state estimation, a critical component in the energy management system of the smart grid [3]–[5].

In addition, there remain significant gaps between the understanding of cybersecurity risks in traditional cyber systems and risks to industrial cyber-physical systems. To bridge the gap, in this paper we study data integrity attacks (i.e., false data injection (FDI) attack) on IIoT systems as a case study. Note that the FDI attack has been a pernicious cybersecurity threat to a number of smart-world systems, such as the smart grid [3], [6], [7], smart manufacturing [8]. To be specific, we first study the FDI attack on the Kalman filter, which is a key component in carrying out state estimation in realistic testbeds (i.e., wireless cyber-physical simulator (WCPS)) for IIoT systems[1]. We then propose and study the evolutionary process of the interactions between the attacker and defender, and their changing strategies with the increase of time and knowledge. We use the WCPS testbed to conduct extensive experiments to emulate the evolutionary process and validate the performance of defensive strategies (e.g., k-nearest neighbors (KNN), 1NN-dynamic time warping (1NN-DTW), and recurrent neural networks (RNNs) with long short-term memory (LSTM) units) against attacks in different stages of evolution. A continuous stirred-tank reactor (CSTR) system as the physical system is modeled in our evaluation.

To study the interactions between the attacker and defender, we leverage the kinetic warfare principles from Sun Tzu in the

[1] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Xu, Hansong; Yu, Wei; Liu, Xing; Griffith, David W.; Golmie, Nada T. "On Data Integrity Attacks against Industrial Internet of Things."
Paper presented at IEEE Cyber Science and Technology Congress (CyberSciTech 2020), Calgary, CA. August 17, 2020 - October 24, 2020.

Art of War and Carl von Clausewitz in the *Principle of War*, apply them to the design of cyber warfare tactics at a strategic level [9], [10]. We apply the principle to design a set of defense strategies by considering the different combinations of three basic elements (i.e., strength, space, and time) that attack and defense strategies can manipulate, all in the context of CPS.

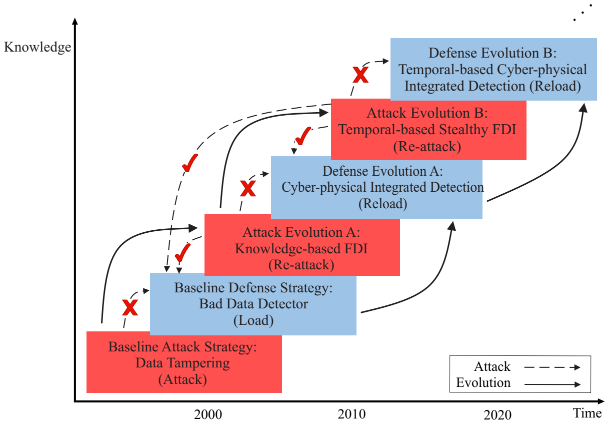

Knowledge
Defense Evolution B: Temporal-based Cyber-physical Integrated Detection (Reload)
Attack Evolution B: Temporal-based Stealthy FDI (Re-attack)
Defense Evolution A: Cyber-physical Integrated Detection (Reload)
Attack Evolution A: Knowledge-based FDI (Re-attack)
Baseline Defense Strategy: Bad Data Detector (Load)
Baseline Attack Strategy: Data Tampering (Attack)
Attack
Evolution
2000
2010
2020
Time

Fig. 1. Overview on the Evolutionary Processes of Attack and Defense Strategies in IIoT

From Fig. 1, the Baseline Attack strategy at the early time stage is a data tampering attack, in which the attacker can modify and inject malicious measurements to affect state estimation and system operations. At the early stage, the attacker launches the data tampering attack by manipulating the strength feature of the system. Accordingly, from the aspect of a defense strategy, the Baseline Defense strategy detects the data tampering attack by investigating the strength feature of the system. In particular, the Baseline Defense strategy can be the x^2 bad data detector (BDD), which ensures the residual of the received measurement and estimated measurement of Kalman filter below a certain threshold [11]. Note that the Kalman filter estimates the system state through the analysis of received measurement data and the system model.

Nonetheless, with an increase in knowledge and time, the attacker can leverage the vulnerability of x^2 BDD to design Attack Evolution A (Knowledge-based FDI (Re-attack)) to bypass the detection [12]. In Attack Evolution A, the attacker manipulates the strength feature of the system with the assistance of system knowledge. Thus, the defense strategy has to evolve to Defense Evolution A (Cyber-Physical Integrated Detection (Reload)), which monitors the behavior of both the physical system and the measurement data. Note that Defense Evolution A conducts detections relying on both the strength (i.e., system residual) and space (i.e., cyberspace and physical space) features of the system.

Moreover, the attacker continues to leverage the vulnerabilities of Defense Evolution A to evolve that attack into Attack Evolution B (Temporal-based Stealthy FDI (Re-attack)), which manipulates not only the strength feature, but also the time feature of the system. In detail, Attack Evolution B is designed to cause limited system deviations that are similar to the system deviations introduced by harsh wireless communication conditions (i.e., heavy noise) intermittently. Eventually, the defensive strategy evolves to Defense Evolution B (Temporal-based Cyber-physical Integrated Detection (Reload)) to defend against Attack Evolution B so that the safety of operations in industrial systems can be ensured. Note that the Defense Evolution B conducts anomaly detections by leveraging both strength, space, and time characteristics of the system.

In short, the attack strategy can evolve rapidly into various forms by combining different system characteristics (i.e., strength, space and time). This makes the design of a solid defense challenging. By following the aforementioned principle, the above defense strategies that we designed, considering the diverse evolutions of an attacker, make the IIoT system unassailable.

To summarize, our contributions are two-fold:

- First, we design the evolutionary process model to capture the natural interactions and evolutions between the attacker and the defender. Particularly, the strategies that either the attacker or defender use to attack or defend the IIoT system evolve as time and knowledge increase.
- Second, we use a realistic IIoT testbed to carry out the investigation of the cybersecurity threats to the IIoT systems. The IIoT testbed collects real data from both the control system and wireless network to be applicable in real-world factories and facilities. We also carry out extensive experiments to validate the effectiveness of the defense strategies in protecting the IIoT system against data integrity attacks.

The remainder of this paper is organized as follows. We review existing research efforts regarding IIoT and cybersecurity in Section II. In Section III, we describe the system model and threat model. In Section IV, we introduce the evolutionary process in detail, which captures the interactions between the attack and defense strategies. In Section V, we present the system implementation and experimental results. We conclude the paper in Section VI.

II. Related Work

Envisioned by Industry 4.0, IIoT has been developed to transform the manufacturing industry from traditional factories to automated smart factories with the integration of advanced information technologies, such as IoT, 5G, machine learning, and others [13]–[15]. Specifically, IIoT is aimed to improve productivity, resource efficiency, and intelligence, and at the same time reduce safety risks, operational overhead and so on through massive interconnections between machinery, sensors, and actuators [16], [17].

There have been a number of research efforts investigating IIoT from a variety of perspectives, including 5G for IIoT [18], data analytics in IIoT [19], and cyberattacks on IIoT [20], among others. For example, Xu *et al.* [15] investigated IIoT from the perspective of a CPS, which integrates the control [21], networking [22], and computing [23] systems as the cyber components of IIoT. Moreover, there have been some research efforts on the co-design of cyber systems, which captures the fundamental relationships of cyber entities to satisfy diverse requirements (e.g., low latency and high reliability), such as in networking and control co-design [24], [25].

Despite the communication advantages gained from massive interconnections, the cybersecurity threats are also raised that could endanger not only the cyber systems, but also the physical critical-infrastructures [1], [8], [26], [27]. There have been several research efforts on the investigation of the cybersecurity issues in IIoT, including the security risks of industrial control systems (ICS) [28], cyberattack paths against critical infrastructures [20], and securing IoT-based critical infrastructures [1].

Uniquely, in this paper, we study data integrity attacks against IIoT from the perspective of the evolutionary process, which naturally captures the interactions between the attacker and the defender. We leverage a realistic testbed for IIoT, which consists of plant, sensors, network, state estimator, and controller, to investigate the interactions between attack and defense strategies. In addition, we collect real data from a temperature control plant (i.e., CSTR) model to conduct analysis. Moreover, we investigate two evolved attack scenarios, i.e., Attack Evolution A and Attack Evolution B, and two evolved defense strategies, i.e., cyber-physical integrated detection (Defense Evolution A) and temporal-based cyber-physical integrated detection (Defense Evolution B). Finally, we carry out extensive experiments to evaluate the performance of the proposed detection strategies in defending against different attack scenarios.

III. System and Threat Model

In this section, we first introduce the system model of the IIoT system, including the key components in the system. We then describe the threat model of FDI attacks, including attack goals and objectives.

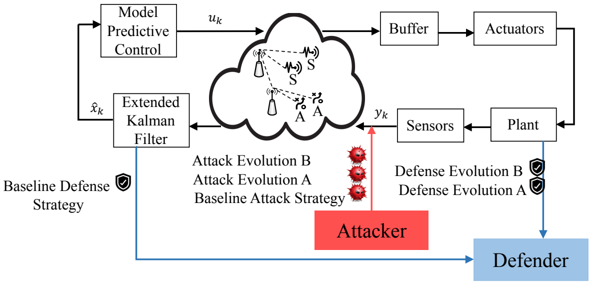


Fig. 2. System Architecture of IIoT with the Integration of Attack and Defense Strategies

We use a linear time-invariant (LTI) system and a wireless network, which transmits sensing data and control commands to form a networked control system. Specifically, we use Fig. 2 to illustrate the IIoT system, which consists of several key components, including the extended Kalman filter (EKF), model predictive control (MPC) controllers, buffer, and wireless networks. As shown in the figure, the attacker injects malicious data into network packets transmitted from sensors and the defender carries out anomaly detection from both cyber and physical aspects.

In particular, the EKF conducts state estimation based on the system model and sensing data received through the wireless networks. Note that, compared to the basic Kalman filter, the EKF conducts state estimation based on not only on the received sensing data, but also on the output of the EKF for prediction. Thus, the EKF is robust to dropped packets of sensing data in wireless networks. We assume the wireless network to be the wireless sensor and actuator network (WSAN) utilizing the WirelessHART protocol, which is a standard communication architecture in IIoT [29].

The MPC controller is the controller for the LTI system. The buffer is used to store the actuation commands to improve the actuation performance to deal with packet drops. Specifically, the MPC controller will generate multiple consecutive control commands and store them in the buffer. In the case of transmission failure of a sensing packet to the controller due to communication noise, the system can use the previously generated actuation commands in the buffer to conduct sub-optimal control.

In this study, we use a classic continuous stirred-tank reactor (CSTR) model as the physical plant in the IIoT system [30]. The CSTR has one main tank for chemical reaction, which requires a certain reaction temperature (i.e., setpoint temperature). This main tank has one inflow to input product, one outflow to output effluent, and a stirrer to mix the products. An additional stream pipe is installed to conduct heat exchange and ensure the main tank has a given setpoint temperature. A valve is equipped on the stream pipe so that the MPC can control the stream rate to adjust the temperature to the desired value. A sensor is equipped on the main tank to measure the tank temperature.

The LTI system, which describes the linearized CSTR, is represented by

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k, \\ y_{k+1} &= Cx_{k+1}, \end{aligned} \tag{1}$$

where x_k represents the state variables at time k, u_k is the control input (e.g., stream rate), w_k represents the system disturbance, y_{k+1} is the system output (e.g., water tank temperature), A and B are Jacobian matrices, and C is the output matrix.

The goal of attack is to mislead the controller to generate wrong control commands and cause deviations to the system without being detected. For example, the attacker can deviate the tank temperature from the setpoint temperature to produce bad chemical products in the CSTR system. We assume that the attacker has full knowledge of the target system and can control some sensors (e.g., the temperature sensor in the CSTR main tank) to inject malicious data to be transmitted to the controller.

To do so, the knowledge-based FDI attack (Attack Evolution A) can carefully design an attack sequence and inject it into network packets, which are distributed by sensors and bypasses the detector (i.e., BDD), as shown in Fig. 2. The attacker has full control over the compromised sensors, in which he or she can inject attack sequences and change the sensing value. The EKF receives the sensing value y_k (containing malicious data) through the wireless network and conducts state estimation as $\hat{x}_k$. Then, the controller generates a sequence of control commands u_k based on the received false sensing data, which could lead to deviation of the system from the setpoint. Note that the evolutionary process of attack strategies (from Attack Evolution A to Attack Evolution B) and defense strategies (from Baseline Defense strategy to Defense Evolution A and further to Defense Evolution B) will be detailed in Section IV.

IV. Evolutionary Process

In this section, we present the evolutionary process of attack and defense strategies.

A. Rationale

The Baseline Attack strategy injects malicious data to tamper with measurement data and impact the system performance. The Baseline Defense strategy computes the square of the errors between the estimated measurement and the received measurement, such that measurements above a threshold will be considered as incorrect measurements and will be removed. The Baseline Defense strategy can identify some arbitrary and intentional bad measurements injected by the attacker effectively. Nonetheless, a sophisticated attack strategy (i.e., Attack Evolution A), which leverages the vulnerability of the Baseline Defense strategy, implements an attack sequence with knowledge of the current system configuration (i.e., CSTR system) to bypass the Baseline Defense strategy [11].

From the defender's perspective, the defense must evolve to Defense Evolution A, checking not only the measurement value, but also the system behaviors. Recall that the attack goal is to cause deviations to the system. The system trajectory captures the system behaviors in a long time window. By comparing normal and malicious system behaviors, Defense Evolution A can identify the existence of malicious data injected by Attack Evolution A.

With the increase in time and knowledge, the attacker will further evolve their strategy to leverage the harsh wireless condition inherent to the industrial environment. The harsh wireless condition, which results in packet drops, will cause system instability and deviations. Thus, Attack Evolution B launches an FDI attack intermittently and in a short time (i.e., 5 to 10 seconds) to mimic the actual harsh wireless condition, causing limited system deviations (similar to the system deviations in harsh wireless condition) repeatedly to bypass the Defense Evolution A.

Furthermore, to defend against Attack Evolution B, the Defense Evolution B (Temporal-based Cyber-physical Integrated Detection), which leverages the memory capability of RNN-LSTM, can identify the existence of Attack Evolution B by leveraging the sequential relationship of system behaviors under attack or harsh wireless conditions.

B. Baseline Attack and Defense Strategies

Recall that the Baseline Attack strategy is to inject malicious measurements to disrupt the system control and achieve malicious goals. It is conceivable that an attacker can compromise the industrial sensors, networking devices, and so on, in the IIoT system to intercept and modify the measurement data. The modified measurement data could produce incorrect state estimations and malicious control commands to the system. In this way, the system could eventually perform poorly, achieving the attacker's goal.

To defend against the Baseline Attack strategy, the Baseline Defense strategy is the x^2 BDD, which detects the error between the estimated sensor measurement and the actual sensor measurement caused by meter failures or malicious attacks. The Baseline Defense strategy will compare the error of the estimated measurement and the actual measurement to a predefined threshold (τ). If the deviation is larger than the threshold, it will infer the presence of bad measurement data. The selection of threshold (τ) is usually based on the hypothesis test with a significant level of (a). This means that the error^2 is larger than threshold^2, inferring the presence of bad measurement data with (a) false alarm rate.

We use Algorithm 1 to carry out the bad data detection. The initialization values are system output y_k, estimated system output $\widehat{y}_k$, as shown in Equation (3)), threshold τ, *DetectionRate*, *Runtime* and *TotalError*. The *DetectionRate* indicates the number of samples that the BDD collected per unit time. The *Runtime* is the time window in which the BDD conducts the detection. The *TotalError* is the threshold determined by the hypothesis test. Then, Algorithm 1 compares z_k to τ in line 6 to determines the existence of bad data in a given set of data samples. The total time complexity of Algorithm 1 is $O(N)$, consisting of $O(N)$ for the while loop and $O(1)$ for the if-else statement. Here, N equals the size of *DetectionRate* × *Runtime*.

Note that we define the residue, which measures the error between the estimated system output $\widehat{y}_k$ and actual measurement y_k as

$$z_k = y_k - \widehat{y}_k, \tag{2}$$

and

$$\widehat{y}_{k+1} = C(A\widehat{x}_k + Bu_k). \tag{3}$$

The state estimation $\widehat{x}_k$ is conducted by the Kalman filter as shown in Equation (4).

Algorithm 1: Detecting Bad Data using x^2 Failure Detector

```
   Result: Alarm Triggered
1  Initialization: $y_k$, $\widehat{y}_k$, $\tau$, DetectionRate, Runtime and TotalError;
2  while $k \leq$ DetectionRate $\times$ Runtime do
3      $z_k = (y_k - \widehat{y}_k)^2$ ;
4  end
5  TotalError $= sum(z_k)$;
6  if TotalError $\leq \tau^2$ then
7      Do nothing;
8  else
9      Trigger False Data Alarm with possibility $P($TotalError $\leq \tau^2)$;
10 end
```

C. Attack Evolution A: Knowledge-based FDI (Re-Attack)

To bypass the Baseline Defense strategy, the attacker can design an attack sequence carefully with the increased system knowledge and inject it into the compromised sensors. We now illustrate how to compute the FDI attack sequence and analyze the FDI attack under the different wireless conditions in the IIoT environment. We assume that the adversary knows the system parameters in Equation (1), as the adversary can be an insider.

The wireless network plays the role of transmitting information between the sensors, actuators, and controllers, such as transmitting the sensed temperature to the state estimator and MPC in CSTR. The wireless network performs differently in a harsh industrial environment, due to channel noise, moving obstacles, etc. [29]. To collect realistic measurements in the industrial environment, Li *et al.* [31] collected the noise traces under wide-range wireless conditions from industrial WSAN testbed and injected them into TOSSIM (a TinyOS simulator) in the WCPS testbed to measure the network performance under different noise levels. The noise levels in the industrial environment ranged from -82 dBm to -72 dBm. The impact on the network performance, measured by link failure rate ranges from 15 % to 98 %, accordingly. Note that the high-level noise usually results from extreme conditions, such as weather or jamming attacks. The results from [31] demonstrate that the EKF and Buffered actuation have very good performance in handling up to 60 % packet drops. Thus, we assume that, under the weak noise levels (e.g., -82 dBm to -75 dBm), the IIoT system operates normally.

We now present the FDI attack on the Kalman filter in the IIoT system under weak wireless conditions. The Kalman filter conducts state estimation based on the received system output. The state estimation is represented in Equation (4) under the assumption that Kalman filter gains K will converge quickly if the system is detectable, as shown in [12]. The state estimation of the Kalman filter is represented as

$$\widehat{x}_{k+1} = A\widehat{x}_k + Bu_k + K(y_k - C(A\widehat{x}_k + Bu_k)), \quad (4)$$

where the $\widehat{x}_{k+1}$ is the estimated state at time $k+1$, and the K represents the Kalman filter gain.

Thus, the system under attack can be represented by

$$y'_k = Cx'_k + v_k + \Gamma R^a_k, \quad (5)$$

where y' and x' are the system output and system state under attack, Γ represents a diagonal matrix to select sensors to manipulate, and R^a_k is the malicious data.

To conduct a "successful" attack, we can leverage Theorem 2 in [12] to compute an attack sequence R^a_k. Specifically, to compute an attack sequence recursively, we have

$$R^a_{k+1} = R^a_k - \frac{\lambda^{i+1}}{M} y^*, \quad (6)$$

where $M = max\,\|\Delta z_k\|$ and $|\lambda| \geq 1$. Note that z_k is the residue defined in Equation (2). Mo *et al.* [12] proved that the system in Equation (1) is "perfectly attackable" if and only if v is the reachable state of the dynamic system (an unstable eigenvector of A) and $Cv \in span(\Gamma)$, in which $span(\Gamma)$ is the column space of Γ. Thus, there exists y^* such that $\Gamma y^* = Cv$.

In addition, heavy noise levels (e.g., -75 dBm to -73 dBm) will cause a link failure rate to above 60 %. Under the heavy noise levels, the control system becomes significantly unstable, experiencing abrupt deviations, as shown in Fig. 5. Note that the EKF and Buffered actuation equipped in the IIoT testbed can eventually bring the control objective of the system to the setpoint. The system deviations under heavy noise levels are significant compared to the system deviations caused by the FDI attack. Thus, the FDI attack becomes ineffective in terms of causing system deviations.

Under extreme wireless conditions (noise levels ranging from -73 dBm to -72 dBm), the packet drop rate will exceed 90 %, which indicates consecutive packet drops. The Kalman filter will receive consecutive zeros in a long time window. The consecutively receiving zeros will make the Kalman filter dysfunctional. Thus, the entire system will go to the fail-safe state to prevent further damage. Moreover, under extreme wireless conditions, the FDI attack is not effective.

To implement Attack Evolution A, we propose Algorithm 2, which computes the malicious system output y'_k and sends it to the EKF via wireless network. Recall that we assume the attacker knows the system configurations. Thus, to launch an FDI attack, the attacker first collects the initial values of system output y_k, estimated system state x_k (follows Equation (4)), residue z_k, M (follows $M = max\,\|\Delta z_k\|$), $\lambda = 1$, v (eigenvectors of A), Kalman filter gain K, y^* (follows $\Gamma y^* = Cv$).

Then, the attacker computes the attack sequence R^a_k (line 3) and injects it into the system (line 4) when the packet drop rate (*PacketDrop*) is below 60 %. Note that the *PacketDrop*, which represents the current wireless conditions statistically, can be computed during a time window. The *AttackLength* determines the duration of the attack. The time complexity of Algorithm 2 is $O(N)$ due to the while loop in the if-else statement, where N equals the size of *AttackLength*.

Algorithm 2: Implementation of Attack Evolution A

```
   Result: malicious system output: $y'_k$
1  Initialization: $y_k$, $x_k$, $M$, $v$, $K$ and $y^*$, $\lambda$, PacketDrop and
     AttackLength;
2  if PacketDrop $\leq 60\%$ then
3      while $k \leq$ AttackLength do
4          $R^a_{k+1} = R^a_k - \frac{\lambda^{i+1}}{M} y^*$;
5          $y'_k = Cx'_k + v_k + \Gamma R^a_k$;
6      end
7  else
8      Do Nothing;
9  end
```

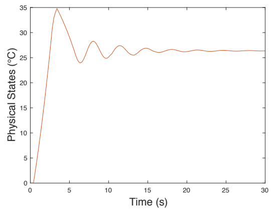


Fig. 3. System Output of a Control System under Normal Conditions

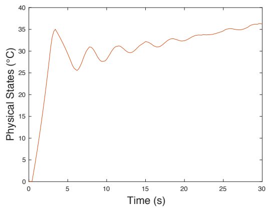


Fig. 4. System Output of a Control System under FDI Attack

As a result, Fig. 3 illustrates the attack-free control system performance, where the MPC controller quickly drives the system to the setpoint (i.e., 27°C for the CSTR system). Fig. 4 shows the system under a continues FDI attack, which leads to significant deviation (i.e., from 27°C to 35°C). Note that both cases do not trigger the alarm of BDD.

D. Defense Evolution A: Cyber-physical Integrated Detection (Reload)

To defend against the knowledge-based FDI (Attack Evolution A), the defender seeks to leverage both the strength and the space characteristics of the system. The detection strategy is implemented in the physical system to collect not only the measurement data, but also the system behaviors. This is because the attack objective of the FDI attack is to disrupt the control system and cause deviations to the physical system. Thus, the Defense Evolution A can leverage a distance-based classifier on the plant (e.g., CSTR water tank) to monitor the output of the LTI system. In addition, to mitigate the deviations caused by the FDI attack on the LTI system, the design goal for the classifier is not only the accurate detection of attacks, but also the rapid detection of attacks (using a small number of data points).

From Fig. 4, the deviations caused by the FDI attack is one significant feature that can be leveraged to design a classifier. This is because the distance of the system behaviors between normal conditions and attack conditions is significant. In this case, we can leverage the K-nearest neighbor (KNN) algorithm as one of the representative classifiers to conduct supervised learning. Specifically, the KNN computes the Euclidean distance of the given input to all training samples and classifies the input as the majority category of the κ nearest samples.

We use Algorithm 3 to conduct the detection of the FDI Attack using the distance-based classifier. The first step is to prepare the data for training and testing, such as data samples under normal conditions *xNormal* and their label *yNormal*, and data samples under FDI attack *xAttack* and their label *yAttack*. Then, for initialization, we separate the data into 70 % training samples (Normal and Attack) and 30 % testing samples: *XTrain* and its label *YTrain*, and *XTest* and its label *YTest*. We also determine the optimal number of κ neighbors by testing the performance of κ from 1 to 10. The total time complexity of Algorithm 3 is $O(N^2)$ due to the two nested loops.

Nonetheless, due to the time-series nature of the LTI signals, when two LTI signals are not aligned, the Euclidean distance will increase, which could interfere with the classification results. To deal with this problem, we use dynamic time warping (DTW) to find the optimal alignment between two time-series signals. Specifically, 1NN-DTW, which classifies the input based on the DTW distance of one nearest neighbor, can achieve state-of-the-art classification performance on time-series classification tasks [32].

Algorithm 3: Detecting FDI Attack using Distance-based Classifier (KNN)

```
  Result: $Results$
1 Data Preparation: xNormal, yNormal, xAttack and
    yAttack;
2 Initialization: XTrain, YTrain, XTest, YTest, $\kappa$;
3 for $i = 1 : length(XTest)$ do
4     for $j = 1 : length(XTrain)$ do
5         Compute EuclideanDistance(XTest(i),
            XTrain(j));
6     end
7     Compute $I(i)$ containing indices for $\kappa$ smallest
        Euclidean distance EuclideanDistance(XTest(i),
        XTrain(j));
8     Assign majority labels: $Results = YTrain(I(i))$
9 end
```

To compute the FDI attack detection using 1NN-DTW, we use Algorithm 3 and update the *EuclideanDistance*(*XTest*(i), *XTrain*(j)) in line 5 with *DTWDistance*(*XTest*(i), *XTrain*(j)). The *DTWDistance* is computed following

$$\textit{DTWDistance}(i,j) = \|x(i) - y(j)\| + \\ min\begin{cases} \textit{DTWDistance}(i, j-1) \\ \textit{DTWDistance}(i-1, j) \\ \textit{DTWDistance}(i-1, j-1) \end{cases}$$

, which combines Euclidean distance $\|x(i) - y(j)\|$ and the cumulative distance of nearest neighbors. In this way, the 1NN-DTW can measure the similarity of two time-series signals more accurately.

E. Attack Evolution B: Temporal-based Stealthy FDI (Re-Attack)

To design a better attack strategy to bypass the cyber-physical integrated detection, the attacker leverages not only the strength characteristic, but also the time characteristic. In other words, the attacker reduces the strength of the FDI attack and diffuses the FDI attack in the time domain.

Specifically, recall that cyber systems, such as networking and control systems, are highly interrelated in the IIoT environment. The communication noise in the networking system also impacts the stability of the control system. Thus, an intelligent attacker could leverage the imperfection nature of the networking system to disrupt the distance-based classifiers. In particular, the attacker can launch the FDI attack intermittently and periodically and only cause deviations that are similar to the deviations caused by packet drops in heavy noise. In this way, the performance of a distance-based classifier will decrease significantly as the classifier would fail to distinguish between system deviations caused by FDI attacks from packet drops under harsh wireless conditions.

We use Fig. 5 to visualize the comparison of the system behaviors under FDI attacks, heavy noise, and attack-free scenarios. Observe that, if the attacker launches periodic FDI attacks of short duration (e.g., 1-10 seconds), the Euclidean distance will no longer be capable of discriminating between malicious and benign signals. This is equivalent to the scenarios that a detector can only use small observation windows to

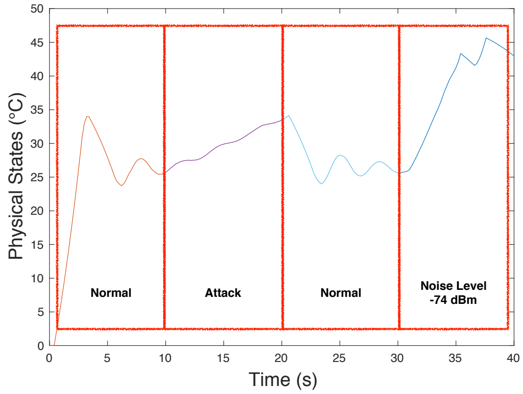


Fig. 5. System Behaviors under Normal, Attack and Heavy Noise Levels

conduct classification. This type of attack over a long duration can cause system deviations that are harmful to industrial systems.

We use Algorithm 4 to implement Attack Evolution B. The initial values are the Attack Evolution B period (AEBP), hiding period (HP), sampling rate (SR) and the initial values in Algorithm 2, such as y_k, x_k, M, v, K and y^*, λ and *PacketDrop*. Specifically, the algorithm repeatedly computes malicious system output y'_k, containing the y'_k with false data R^a_k injected during AEBP and unchanged system output $y'_k = Cx'_k + v_k$ during HP. Note that the SR determines the number of data points per unit time(s). The total time complexity for Algorithm 4 is $O(N)$, consisting of $O(N)$ for the first loop and $O(N)$ for the second loop. Here, N equals to the size of $AEBP \times SR$.

Algorithm 4: Implementation of Attack Evolution B

```
Result: malicious system output: $y'_k$
Initialization: AEBP, HP, SR, $y_k$, $x_k$, $M$, $v$, $K$ and $y^*$,
  $\lambda$ and PacketDrop;
repeat
    for $k = 1 : (AEBP \times SR)$ do
        if PacketDrop $\leq 60\%$ then
            $R^a_{k+1} = R^a_k - \frac{\lambda^{i+1}}{M} y^*$;
            $y'_k = Cx'_k + v_k + \Gamma R^a_k$;
        else
            Do Nothing;
        end
    end
    for $k = (AEBP \times SR) :$
      $(AEBP \times SR + HP \times SR)$ do
        $R^a_k$ = 0 ;
        $y'_k = Cx'_k + v_k + \Gamma R^a_k$;
    end
until STOP;
```

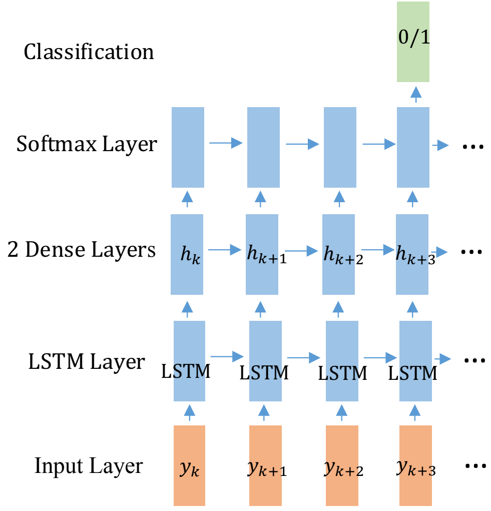


Fig. 6. RNN Architecture

F. Defense Evolution B: Temporal-based Cyber-Physical Integrated Detection (Reload)

Defense Evolution A fails to distinguish between system behaviors under Attack Evolution B and heavy noise levels in the wireless network. This is because the distance between the system deviations under heavy noise and Attack Evolution B declines. Thus, the classification accuracy of distance-based classifiers decreases and at the same time, more data points are necessary to conduct classification accurately. To overcome the issue, we design the temporal-based cyber-physical integrated detection, which leverages not only the strength characteristic, but space and time characteristics to conduct anomaly detection.

In detail, we leverage the recurrent neural network, which has proven effective on natural language processing (NLP) tasks, such as voice recognition, in addition to many others [33], [34]. Due to its memorization capability, the RNN-LSTM can be used to solve sequence classification problems. This is because the RNN-LSTM accumulates information extracted from previous observations and adjusts the next decision based on the historical observations. Thus, RNN-LSTM can provide better classification accuracy in terms of sequential data classification. In particular, we use the many-to-one architecture, as shown in Fig. 6, which takes the sequential control signal as input, examines this sequence sample-by-sample, and gives a decision at the last sample. The RNN-LSTM follows a standard sequence-to-label classification design, which contains a sequence input layer, an LSTM layer, a dense (fully connected) layer, a softmax layer, and a classification output layer.

V. Experimental Results

We have implemented the attack and defense strategies outlined above in an IIoT testbed and have conducted extensive experiments to study the evolutionary process. We used the classic CSTR in MATLAB as our physical system [30]. The network settings followed the WirelessHART network topology to simulate a WSAN. The testbed adopted TOSSIM to simulate the WSAN with adjustable noise levels.

We implemented the 1NN-DTW and KNN (Defense Evolution A) in MATLAB, which takes the LTI system as input and outputs the classification result (benign or malicious). We collected 1,000 samples of attack-free LTI signals and 1,000 samples of LTI signals under attack. Additionally, we used 70 % of the samples to train the 1NN-DTW and KNN, and the remaining 30 % of samples were used for testing. We also implemented the RNN-LSTM based classifier of Defense Evolution B in MATLAB. The RNN-LSTM architecture consists of an input layer, an LSTM layer, a fully connected layer, a softmax layer, and an output layer. We used the back-propagation through time (BPTT) as the gradient descent algorithm. Similarly, we used a 70 %/30 % sample split for training/testing the RNN-LSTM.

The classification performance is quantized from the aspect of both accuracy and timeliness. The accuracy is measured by the percentage of correctly identified time-series data samples over the total number of testing samples. The timeliness means how many data points have been used to obtain the classification accuracy. The larger the number of data points used, the longer it took to conduct classification, assuming that the sampling rate is fixed.

A. *Performance of Defense Evolution A in Defending Attack Evolution A*

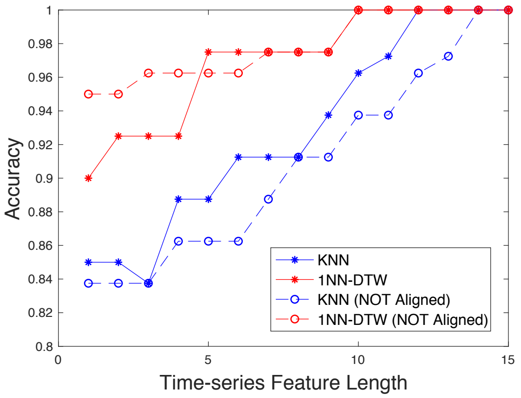


Fig. 7. Classification Performance of KNN vs 1NN-DTW

We now illustrate the experimental performance of the Defense Evolution A (i.e., distance-based anomaly detection (reload)) vs. the Attack Evolution A (i.e., knowledge-based FDI (re-attack)). Observe from Fig. 7 that the 1NN-DTW can obtain nearly 97.5 % in classification accuracy at a single control loop (5 data points per second at a sampling rate equal to 5 HZ) and 100 % classification accuracy at two control loops (10 data points in 2 seconds). The KNN can obtain 88.8 % classification accuracy at a single control loop, 96.3 % classification accuracy at two control loops, and 100 % classification accuracy at three control loops (15 data points in 3 seconds). Note that the ’NOT Aligned’ signifier indicates the signals have not been manually aligned. The red lines show the performance of 1NN-DTW and KNN without aligning the signals manually. Benefitting from the DTW technique, the 1NN-DTW performs better when the signals are not manually aligned.

B. *Performance of Defense Evolution A and Defense Evolution B in Defending against Attack Evolution B*

We further illustrate the experimental performance of Defense Evolution A and Defense Evolution B vs. Attack Evolution B under different wireless conditions.

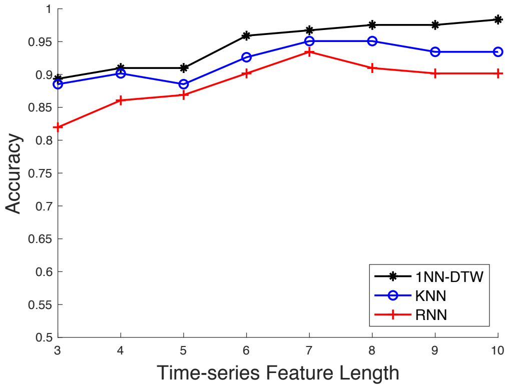


Fig. 8. Detect Attack Evolution B under Weak Noise Levels

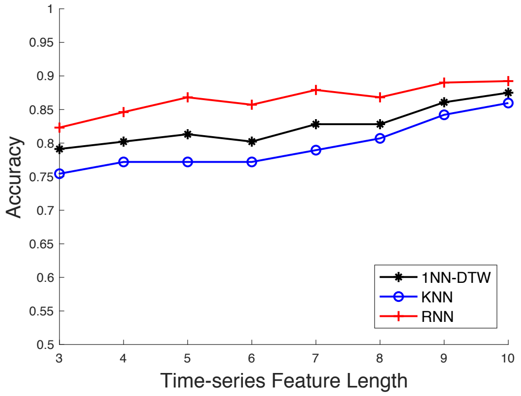


Fig. 9. Detect Attack Evolution B under Weak and Heavy Noise Levels

Observing from Fig. 8, we can see that both 1NN-DTW and KNN outperform RNN-LSTM under a weak communication noise environment (Defense Evolution A defend against Attack Evolution A). Specifically, the detection accuracy of 1NN-DTW increases from nearly 89.3 % with only three data points to 98.4 % with 10 data points. In addition, the detection accuracy of 1NN-DTW increases faster than the others when the number of data points used increases. This is due to the signal alignment capability of the DTW.

The RNN-LSTM shows the lowest classification accuracy in comparison to the distance-based classifiers. This is because the time-domain sequential feature is not significant compared to the distance feature. In contrast, Fig. 9 shows the inadequacy of Defense Evolution A in defending against Attack Evolution B, where the distance feature used by Defense Evolution A has been disrupted by the heavy noise. The system deviations caused by the packet drops under weak and heavy noise levels misleads the distance-based classifiers. Thus, the detection accuracy of KNN and 1NN-DTW drop to 75.4 % and 79.1 % with 3 data points, respectively. With 10 data points, the accuracies are 85.9 % and 87.5 %. In comparison, the RNN obtains a higher classification accuracy, which increases from 82.3 % with 3 data points to 89.2 % with 10 data points.

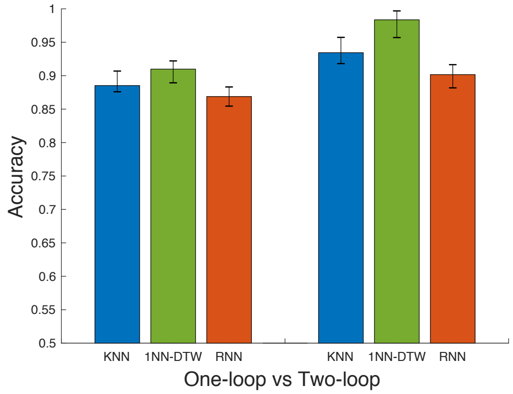


Fig. 10. Classification Performance per Control-Loops under Weak Noise Levels

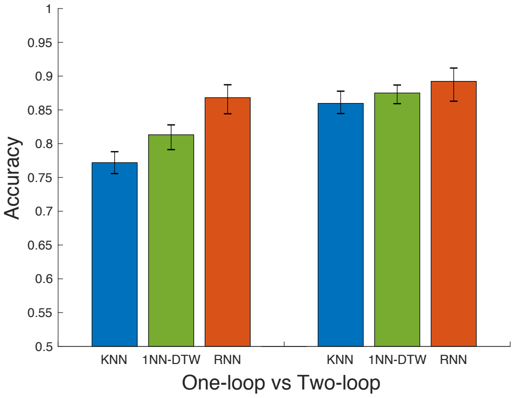


Fig. 11. Classification Performance per Control-Loops under Weak and Heavy Noise Levels

Recall that, as the goal of the defender is to detect the FDI accurately and rapidly, we compare the classification accuracy with regard to detection speed, as shown in Fig. 10 and Fig. 11. Observing from Fig. 10, we can see that the 1NN-DTW outperforms KNN and RNN-LSTM and achieves an average classification accuracy of 90.9 % and 98.3 % at one-loop and two-loop under the weak noise scenario. In contrast, from Fig. 11, the RNN-LSTM outperforms 1NN-DTW and KNN and obtains a classification accuracy of 86.8 % and 89.2 % at one-loop and two-loop under weak and heavy noise levels.

In addition, we use the receiver operating characteristic (ROC) curve to measure the performance of classifiers in Fig. 12 and Fig. 13 with the false positive rate and true positive rate on the x-axis and y-axis, respectively. Note that we use the area under the curve (AUC) to quantitatively measure the separability of classifiers between 0.5 to 1 (the higher the better). Notably, the 1NN-DTW obtains 0.98 of AUC in weak noise levels, the highest when compared to other methods. Additionally, the RNN-LSTM obtains 0.90 of AUC in weak and heavy noise levels, the highest compared to the other classifiers.

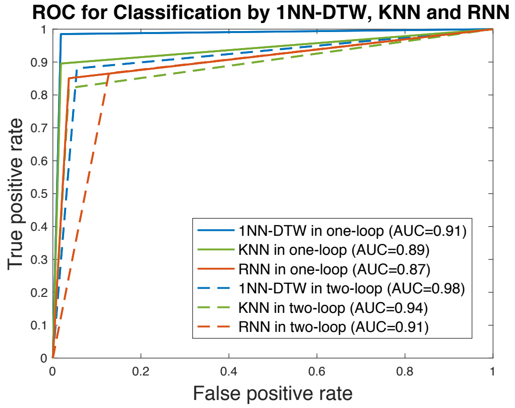


Fig. 12. ROC Curve for Classifiers under Weak Noise Levels

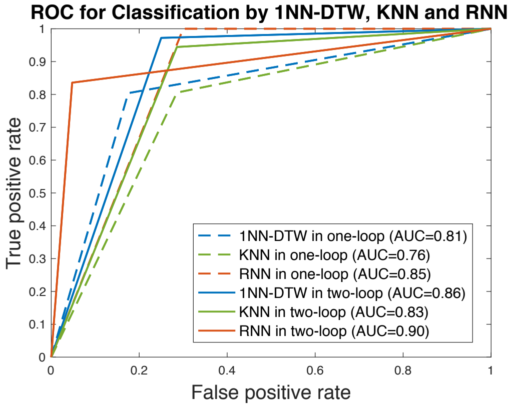


Fig. 13. ROC Curve for Classifiers under Weak and Heavy Noise Levels

VI. Final Remark

In this paper, we have studied the evolutionary process of the strategies that an attacker uses to launch data integrity attacks against IIoT and the strategies that a defender uses to detect such attacks, as knowledge and time increase. We have leveraged a realistic IIoT testbed to implement multiple attack and defensive strategies, and have evaluated their interactions. We have also conducted extensive experiments to validate the effectiveness of the defensive strategies against increasingly complex data integrity attacks under different levels communication noise. The experimental results demonstrate strong classification accuracy for 1NN-DTW under weak communication noise and RNN-LSTM under weak and heavy communication noise in wireless communication.

References

[1] X. Liu, C. Qian, W. G. Hatcher, H. Xu, W. Liao, and W. Yu, "Secure internet of things (iot)-based smart-world critical infrastructures: Survey, case study and research opportunities," *IEEE Access*, 2019.

[2] A. Nicholson, S. Webber, S. Dyer, T. Patel, and H. Janicke, "Scada security in the light of cyber-warfare," *Computers & Security*, vol. 31, no. 4, pp. 418 – 436, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167404812000429

[3] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 13:1–13:33, Jun. 2011. [Online]. Available: http://doi.acm.org/10.1145/1952982.1952995

[4] Q. Yang, J. Yang, W. Yu, D. An, N. Zhang, and W. Zhao, "On false data-injection attacks against power system state estimation: Modeling and countermeasures," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 717–729, 2013.

[5] J. Lin, W. Yu, X. Yang, G. Xu, and W. Zhao, "On false data injection attacks against distributed energy routing in smart grid," in *Proceedings of the 2012 IEEE/ACM Third International Conference on Cyber-Physical Systems*. IEEE Computer Society, 2012, pp. 183–192.

[6] B. Li, R. Lu, W. Wang, and K.-K. R. Choo, "Distributed host-based collaborative detection for false data injection attacks in smart grid cyber-physical system," *Journal of Parallel and Distributed Computing*, vol. 103, pp. 32–41, 2017.

[7] D. Ding, Q.-L. Han, Y. Xiang, X. Ge, and X.-M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[8] D. Wu, A. Ren, W. Zhang, F. Fan, P. Liu, X. Fu, and J. Terpenny, "Cybersecurity for digital manufacturing," *Journal of Manufacturing Systems*, vol. 48, pp. 3 – 12, 2018, special Issue on Smart Manufacturing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278612518300396

[9] S. Tzu, "The art of war," in *Strategic Studies*. Routledge, 2014, pp. 86–110.

[10] C. Clausewitz and F. N. Maude, *On war*. Penguin UK, 1982.

[11] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Transactions on Information and System Security (TISSEC)*, vol. 14, no. 1, p. 13, 2011.

[12] Y. Mo and B. Sinopoli, "False data injection attacks in control systems," in *Preprints of the 1st workshop on Secure Control Systems*, 2010, pp. 1–6.

[13] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE industrial electronics magazine*, vol. 11, no. 1, pp. 17–27, 2017.

[14] K. Wang, Y. Wang, Y. Sun, S. Guo, and J. Wu, "Green industrial internet of things architecture: An energy-efficient perspective," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 48–54, 2016.

[15] H. Xu, W. Yu, D. Griffith, and N. Golmie, "A survey on industrial internet of things: A cyber-physical systems perspective," *IEEE Access*, vol. 6, pp. 78 238–78 259, 2018.

[16] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, "A survey on internet of things from industrial market perspective," *IEEE Access*, vol. 2, pp. 1660–1679, 2014.

[17] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: Challenges, opportunities, and directions," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724–4734, 2018.

[18] J. Navarro-Ortiz, S. Sendra, P. Ameigeiras, and J. M. Lopez-Soler, "Integration of lorawan and 4g/5g for the industrial internet of things," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 60–67, 2018.

[19] M. H. ur Rehman, E. Ahmed, I. Yaqoob, I. A. T. Hashem, M. Imran, and S. Ahmad, "Big data analytics in industrial iot using a concentric computing model," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 37–43, 2018.

[20] I. Stellios, P. Kotzanikolaou, M. Psarakis, C. Alcaraz, and J. Lopez, "A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3453–3495, 2018.

[21] T. Wang, D. Wang, and K. Wu, "Chaotic adaptive synchronization control and application in chaotic secure communication for industrial internet of things," *IEEE Access*, vol. 6, pp. 8584–8590, 2018.

[22] X. Li, D. Li, J. Wan, C. Liu, and M. Imran, "Adaptive transmission optimization in sdn-based industrial internet of things with edge computing," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1351–1360, 2018.

[23] M. Aazam, S. Zeadally, and K. A. Harras, "Deploying fog computing in industrial internet of things and industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4674–4682, 2018.

[24] G. Zhao, M. A. Imran, Z. Pang, Z. Chen, and L. Li, "Toward real-time control in future wireless networks: communication-control co-design," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 138–144, 2018.

[25] H. Seo, J. Park, M. Bennis, and W. Choi, "Communication and consensus co-design for low-latency and reliable industrial iot systems," *arXiv preprint arXiv:1907.08116*, 2019.

[26] Z. Szabó, "Cybersecurity issues in industrial control systems," in *2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 2018, pp. 000 231–000 234.

[27] B. Huang, A. A. Cardenas, and R. Baldick, "Not everything is dark and gloomy: Power grid protections against iot demand attacks," in *28th USENIX Security Symposium (USENIX Security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 1115–1132. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/huang

[28] L. A. Maglaras, K.-H. Kim, H. Janicke, M. A. Ferrag, S. Rallis, P. Fragkou, A. Maglaras, and T. J. Cruz, "Cyber security of critical infrastructures," *ICT Express*, vol. 4, no. 1, pp. 42–45, 2018.

[29] C. Lu, A. Saifullah, B. Li, M. Sha, H. Gonzalez, D. Gunatilaka, C. Wu, L. Nie, and Y. Chen, "Real-time wireless sensor-actuator networks for industrial cyber-physical systems," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 1013–1024, 2015.

[30] N. Kamala, "Studies in modeling and design of controllers for a nonideal continuous stirred tank reactor," 2013.

[31] B. Li, Y. Ma, T. Westenbroek, C. Wu, H. Gonzalez, and C. Lu, "Wireless routing and control: a cyber-physical case study," in *2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE, 2016, pp. 1–10.

[32] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *2017 international joint conference on neural networks (IJCNN)*. IEEE, 2017, pp. 1578–1585.

[33] W. De Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61–98, 2015.

[34] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24 411–24 432, 2018.

Automatic Volumetric Segmentation of Additive Manufacturing Defects with 3D U-Net

Vivian Wen Hui Wong,[1] Max Ferguson,[1] Kincho H. Law,[1] Yung-Tsun Tina Lee,[2] Paul Witherell[2]

[1]Engineering Informatics Group, Civil and Environmental Engineering, Stanford University, Stanford, CA 94305

[2]Systems Integration Division, National Institute of Standards and Technology (NIST), Gaithersburg, MD 20899

[1]{vwwong3, maxferg, law}@stanford.edu; [2]{yung-tsun.lee, paul.witherell}@nist.gov

Abstract

Segmentation of additive manufacturing (AM) defects in X-ray Computed Tomography (XCT) images is challenging, due to the poor contrast, small sizes and variation in appearance of defects. Automatic segmentation can, however, provide quality control for additive manufacturing. Over recent years, three-dimensional convolutional neural networks (3D CNNs) have performed well in the volumetric segmentation of medical images. In this work, we leverage techniques from the medical imaging domain and propose training a 3D U-Net model to automatically segment defects in XCT images of AM samples. This work not only contributes to the use of machine learning for AM defect detection but also demonstrates for the first time 3D volumetric segmentation in AM. We train and test with three variants of the 3D U-Net on an AM dataset, achieving a mean intersection of union (IOU) value of 88.4%.

1 Introduction

Additive manufacturing (AM), also known as three-dimensional (3D) printing, is experiencing rapid adoption in the manufacturing domain. AM introduces several advantages such as the fabrication of models with complex geometries, quick prototyping, customization of materials and flexibility in design (Ngo et al., 2018). During an AM fabrication process, consecutive layers are printed (Gross et al., 2014). Internal defects can be created due to reasons such as print error, residual stress, or cyber-attack (Holzmond and Li, 2017). The presence of defects leads to flaws, such as insufficient material properties, in the printed object (Reese, Bheda and Mondesir, 2016). An automatic defect segmentation system to identify interior defects can therefore aid AM quality control.

Some current non-destructive defect segmentation techniques include infrared radiation monitoring and X-ray computed tomography (XCT). The former monitors radiation given off by melt pools during the printing process. The monitoring process requires a complex laser setup and only works for the powder bed fusion (PBF) process (Holzmond and Li, 2017). XCT can be used to visualize internal structures, including porosity, in three dimensions (Buffiere et al., 2001; Masad et al., 2002). While XCT can be used to obtain images and segmentation labels, it requires manual thresholding, which can be tedious given many samples. Other approaches including analyzing an array of sensor signals to monitor the process (Rao et al., 2015), but such approaches require the installment of multiple sensors and analysis of multiple signal types. Automatically identifying small defects inside an object therefore remains a challenging task in manufacturing.

Defect segmentation can be treated as an image segmentation problem in computer vision. In defect segmentation, each 2D pixel or 3D voxel is classified as either defect or background. Image segmentation is a difficult task, and many methods have been proposed to solve the problem, especially in the medical imaging domain, which requires localization of objects (Ronneberger, Fischer and Brox, 2015). State of the art 2D and 3D segmentation methods that perform best in various segmentation tasks are based on deep learning and use convolutional neural networks (CNN) (Guo et al., 2018). The benefits and the drawbacks of some of the current segmentation methods are further discussed in Section 2.

Despite their drawbacks, 3D CNNs have demonstrated potential in 3D medical image segmentation tasks (Cicek et al., 2016; Milletari, Navab and Ahmadi, 2016; Lee et al., 2017; Yu et al., 2017; Zhou et al., 2018; Ghavami et al., 2019). AM images share similar characteristics with medical images such as their volumetric nature and similar level of contrast. Therefore, it is thought that certain deep learning approaches from the biomedical domain might be of similar benefit in the AM domain. However, processing XCT data from AM parts poses certain difficulties that are less prevalent in the biomedical domain, such as small and highly irregular defect geometries. In addition, the sparsity

Wong, Vivian; Ferguson, Max; Law, Kincho; Lee, Yung-Tsun; Witherell, Paul. "Automatic Volumetric Segmentation of Additive Manufacturing Defects with 3D U-Net." Paper presented at SSS20 (VIRTUAL) AAAI Spring Symposium on AI in Manufacturing, Palo Alto, CA, US. March 23, 2020 - March 25, 2020.

of defects in AM varies dramatically, as exemplified by the dataset used in this paper which contains samples that range from 0.37% to 19.38% porosity. Furthermore, AM datasets are both difficult and costly to produce and hence there are very few publicly available datasets large enough to employ machine-learning approaches. Despite these challenges, developing fast and reliable quality control procedures is essential to the mainstream adoption of additive manufacturing processes.

The need for automatic manufacturing defect detection and the promising performance of CNN in the volumetric segmentation of medical images thus motivates our work. We propose to train a 3D U-Net model with existing defect labels and use the trained model to automate the segmentation process on XCT or 3D images of unknown additive manufacturing samples. To demonstrate the effectiveness of our method, an experiment is conducted using an AM defect dataset constructed with XCT images. We train and evaluate 3 variants of the 3D U-Net model and obtain a mean intersection of union (IOU) value of 0.863 to 0.884. Although no previous work has been done in AM defect segmentation to compare this accuracy with, our mean IOU is comparable with the accuracy of 3D U-Net segmentation of kidney embryos, which achieves a mean IOU of 0.704 (Cicek et al., 2016).

The rest of the paper is organized as follows. Section 2 provides an overview of related works. Section 3 briefly introduces 3D CNNs and presents the network architecture of our 3D U-Net models. Section 4 discusses our implementation and experimental results. Finally, Section 5 briefly concludes our work.

2 Related Works

This section reviews current advancements in 3D image segmentation using deep learning approaches, as well as related works in automatic AM defect identification.

Segmentation of 3D objects are often framed as a 2D segmentation task with post processing (Zhou et al., 2016; Milletari, Navab and Ahmadi, 2016). The most commonly used CNN architectures used for 2D segmentation problem are region-based and fully-convolutional-network-based (FCN-based) (Guo et al., 2018).

Region-based segmentation methods first extract regions and describe them, then classify the region (Caesar, Uijlings and Ferrari, 2016). Region-based CNN (R-CNN) is a representative method using this approach (Guo et al., 2018). The Mask R-CNN architecture is an example of R-CNN that performs object detection and segmentation simultaneously (He et al., 2017). Mask R-CNN has proven to be effective in segmenting everyday objects (He et al., 2017), medical images (Johnson, 2018) and metal casting defects in manufacturing (Ferguson et al., 2018).

On the other hand, FCN-based methods perform segmentation by directly learning a mapping from input to output pixels, without proposing regions (Long, Shelhamer and Darrell, 2015). U-Net is a CNN model that extends the FCN architecture, achieving excellent performance in the segmentation of ventral nerve chord (Ronneberger, Fischer and Brox, 2015). U-Net is less computationally expensive than Mask R-CNN, since it does not require the generation of region proposals.

Despite the success of the above-mentioned methods, AM specimens are 3D, and therefore predicting one 2D slice at a time loses information on the correlation between slices (Yu et al., 2019). A similar problem exists in the medical imaging domain, where most images are 3D volumes. To that end, 3D CNN models have been developed to make predictions on volumetric medical images (Cicek et al., 2016; Milletari, Navab and Ahmadi, 2016).

While 3D CNN models can leverage information between slices, they lack pre-trained models, leading to less stable training (Yu et al., 2019). Patch-wise predictions are also more time-consuming to generate compared to 2D predictions. V-Net (Milletari, Navab and Ahmadi, 2016) and 3D U-Net (Cicek et al., 2016) are examples of end-to-end architectures for 3D segmentation. In this work, we adopt 3D U-Net and two of its variants in our experiments with AM defect data.

A few related works have been done to automatically detect AM defects using CNN. Scime and Beuth (2018) and Zhang, Liu and Shin (2019) perform 2D detection and classification on defects using slices of camera images in their work. Shevchik et al. (2018) use CNN to analyze acoustic emissions during AM processes.

As discussed in this section, the use of 3D CNN to segment AM defects represents a new and novel approach, and will be the subject of discussion in the rest of this paper.

3 3D Convolutional Neural Networks

In this section, we present some background knowledge on 3D CNNs. As discussed in the previous section, 2D CNNs and 3D CNNs each have their own drawbacks and advantages. 3D CNNs extend upon 2D CNNs by using the same technique of convoluting a kernel spatially through the input of a convolutional layer with two important distinctions:

1) The kernel of a 2D convolutional layer is two dimensional (2D) with width and height (W × H), and the kernel of a 3D convolutional layer is three dimensional (3D) with width, height and depth (W × H × D).
2) A 2D kernel moves in 2 directions, along the axis corresponding to W and H dimensions. A 3D kernel moves in 3 directions, with an additional axis along the D dimension.

3D U-Net

The 3D U-Net is an extension upon the standard (2D) U-Net architecture proposed by Ronneberger, Fischer and Brox,

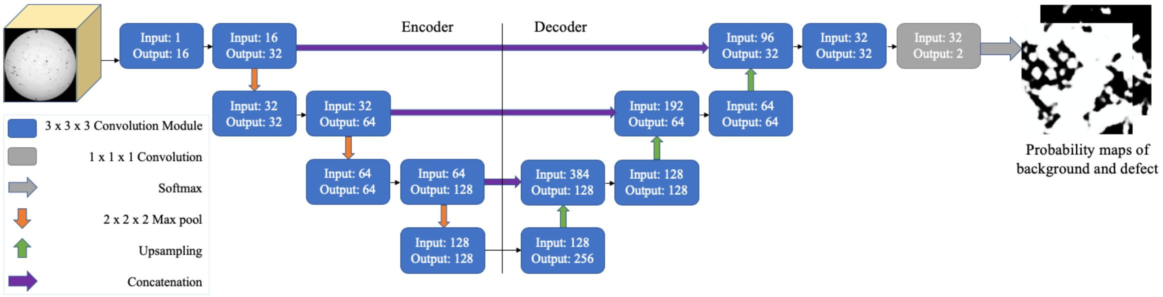


Figure 1: 3D U-Net architecture implemented in this study. The numbers on the convolutional blocks indicate the number of input and output channels (Cicek et al., 2016).

Table 1: Details of the AM defect data

Specimen	Distance between 2D slices [pixel]	Total number of voxels	Porosity (%)	Dataset
Sample 1	0.00245	6.57×10^8	1.01	Validation
Sample 2	0.00277	6.61×10^8	19.28	Training
Sample 3	0.00243	6.55×10^8	0.37	Training
Sample 4	0.00252	5.50×10^8	11.01	Training

(2015). The 3D U-Net can be implemented using a modular architecture, as shown in Figure 1. The architecture consists of two types of modules, encoder modules shown on the left and decoder modules on the right. The encoder modules perform max pool operations and the number of feature maps produced increases as the number of layers increases. On the other hand, the decoder modules perform upsampling and the number of feature maps produced decreases as the number of layers increases. This overall encoder-decoder design is preserved for all 3D U-Net models, but the design of each module can be altered. Each 3 x 3 x 3 convolution module consists of a convolutional layer as well as group normalization (GN) or batch normalization (BN) layer, and a rectified linear unit (ReLU) layer (Cicek et al., 2016).

Both GN and BN are techniques to improve the speed and stability of training neural networks (Ioffe and Szegedy, 2015). Wu and He (2018) propose GN, which is more robust than BN, and does not suffer the limitation of BN that smaller batch size leads to larger errors. The 3D U-Net in the work by Cicek et al. (2016) uses BN and reports outstanding results. To conduct the experiments on the AM defects, the 3D U-Net implementation by Cicek et al. (2016) is employed. However, we modify the convolutional modules to replace BN with GN, as well as rearranging the layer structures, allowing us to achieve a higher accuracy on the AM defects, as discussed in the next section.

Residual Symmetric 3D U-Net is an improved variant of 3D U-Net, proposed by Lee et al. (2017). Several changes are made compared to the 3D U-Net architecture, including the addition of a layer by redesigning convolutional modules to include residual skip connections and symmetricity, as well as modifying the downsampling and upsampling techniques. Lee et al. (2017) report that their model exceeds human accuracy in an experiment segmenting neurites in electron microscopic (EM) brain images.

As discussed in the next section, we modify the convolution module of the 3D U-Net architecture to analyze their impact on the model's performance on the AM defect segmentation task. We also compare Residual Symmetric 3D U-Net against the two designs.

4 Implementation Details and Experimental Results

This section describes the implementation and results of the experiments conducted using a dataset of AM defect images.

4.1 Data

The dataset consists of four cylindrical AM specimens, shown in Table 1. The datasets were introduced and analyzed by Kim et al. (2017), and the data are publicly available (Kim et al., 2019). The artificial defects were produced by changing AM processing parameters. The sample contains AM defects due to unoptimized AM processing parameters. Each specimen consists of 8-bit grayscale images of 2D slices. These images are 16-bit raw images obtained using XCT reconstruction processed by adding a median 3D filter and a non-local means filter (Buades, Coll and Morel,

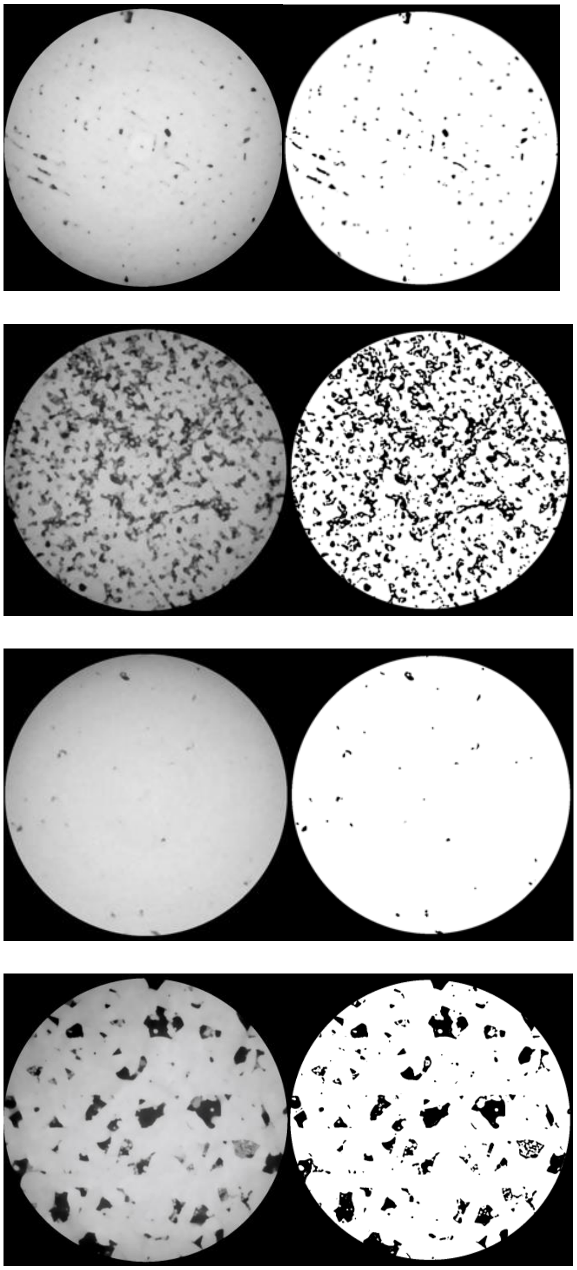


Figure 2: Examples of images from the AM defect dataset. Processed XCT images are on the left and segmentation masks are on the right.

2011; Sun, Brown and Leach, 2012). To obtain the segmentation mask, the 8-bit images are processed with Bernsen local thresholding (Bernsen, 1986). The local contrast threshold parameters of the thresholding process are set by relating average noise value to local contrast threshold as explained by Kim et al. (2017). Examples of some images and corresponding masks are shown in Figure 2.

Table 2: Mean IOU and average training time

Model	Training Time on GPU [hours]	Validation Mean IOU
3D U-Net with Conv+BN+ReLU	**6.58**	0.863
3D U-Net with Conv+ReLU+GN	14.00	0.881
Residual Symmetric 3D U-Net	19.97	**0.884**

In the experimental study, Sample 1 is used for validation. Samples 2, 3 and 4 are used for training. This choice is due to the fact that Sample 1, among all samples, does not have an extreme value of porosity.

4.2 Training and Inference

To analyze the effect of 3D U-Net convolutional modules on performance and convergence, three 3D U-Net models are trained and evaluated in this study. The three models vary in the layers of the basic convolutional modules. The first model, hereby referred to as 3D U-Net with Conv+BN+ReLU, follows the implementation by Cicek et al. (2016). The model uses a 3D convolutional layer, a BN layer and a ReLU nonlinearity layer in its basic convolutional module. The second model, 3D U-Net with Conv+ReLU+GN, is a variant that uses a 3D convolutional layer, a ReLU nonlinearity layer, followed by a GN layer. The third model, Residual Symmetric 3D U-Net, follows the implementation by Lee et al. (2017).

Network inputs are 3D images of dimensions Depth × Width × Height, constructed by stacking the 2D slices, as shown in Figure 3. The inputs are normalized, randomly flipped and rotated prior to training. Network outputs and targets are compared using the softmax function with cross-entropy loss. The models are trained end-to-end and without pretraining for 2000 iterations using a NVIDIA Tesla T4 GPU, which fits patches of size 128 × 128 × 128.

The models are fine-tuned on the AM defect dataset. Each model is trained with an initial learning rate of 0.0002 that decays at a rate of 0.5 at the 600th, 1000th, and 1400th iteration. The networks are trained via the Adam optimizer (Kingma and Ba, 2014). A weight decay factor of 0.0001 is used. We set the batch size and the group size of one in BN and GN, respectively. Stride sizes are 32 × 32 × 32 to overlap the patches and to capture the fine details in the images. All modifications and fine-tuning to the models are conducted using a publicly available implementation of the 3D U-Net architecture (Wolny, 2019).

The models are evaluated with the same GPU. The prediction accuracy of each model is evaluated using the mean intersection over union (IOU) metric.

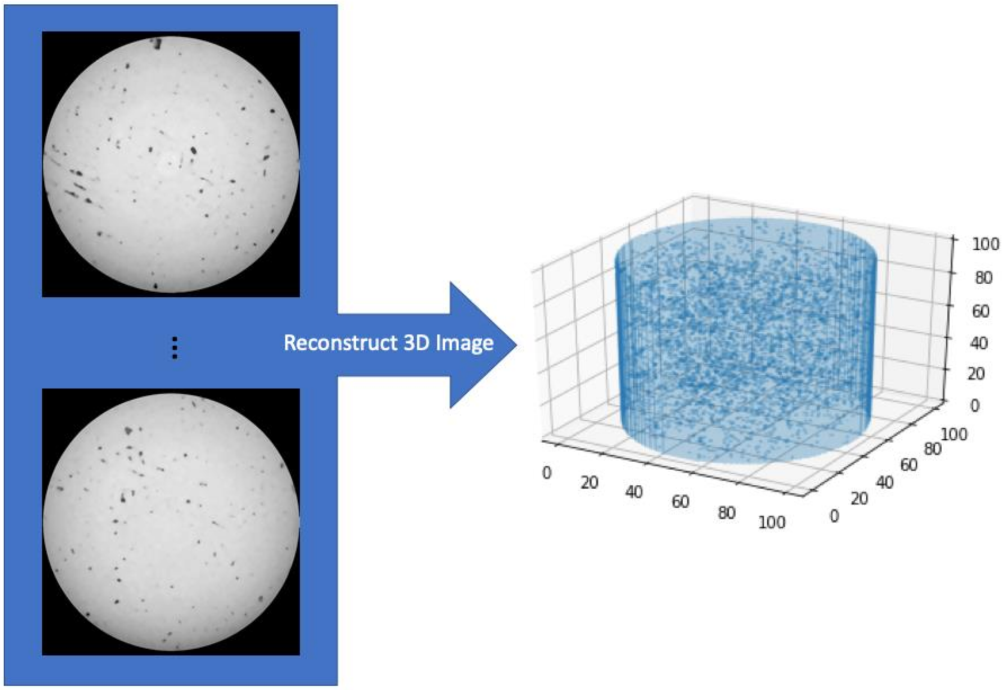


Figure 3: Reconstruction of a 3D image. Note that the voxels on the right are downsized for display purpose.

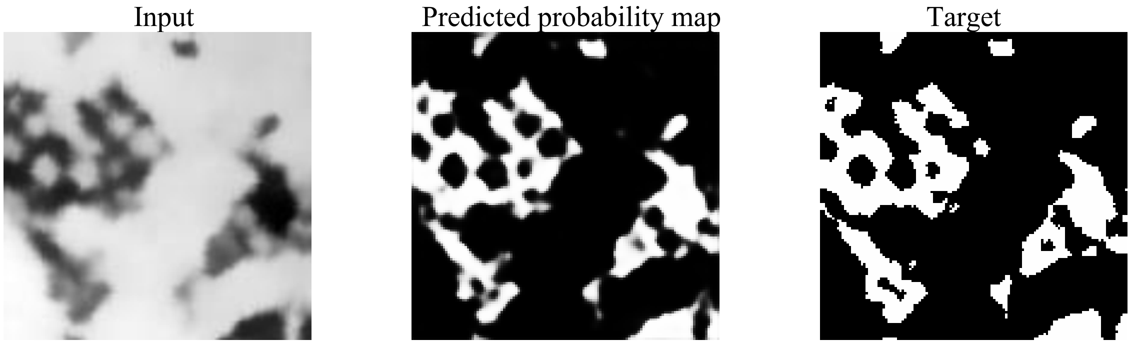


Figure 4: An example slice of a defect and its segmentation probability map outputted by Residual Symmetric 3D U-Net.

4.3 Results and Discussion

The performance of the three different 3D U-Net models are compared in Table 2. Each model's validation accuracy and the amount of time taken to achieve that are reported in the table. The Residual Symmetric 3D U-Net model, with a mean IOU of 0.884, exceeds the other models slightly in performance, but requires the longest training time. With the 3D U-Net model, GN improves the result with respect to BN, but is slower in the training iterations. In general, we observe a trade-off between training time and accuracy. In practice, an appropriate model should be selected taking into consideration the time and accuracy trade-off.

Figure 4 shows an example of a slice segmented using the Residual Symmetric 3D U-Net model. It can be seen that the segmentation probability map compares well with the target (labelled) sample. One observation is that the voxels corresponding to sharp geometries of the defects are often misclassified. Due to the size of the defects and poor contrasts at the edges, these sharp geometries appear to be small and light in color, and their true label is often ambiguous, hence posing difficulties for segmentation.

5 Summary and Discussion

This paper has presented a method for automatic volumetric segmentation of AM specimens using 3D U-Net, a CNN model previously developed for medical image segmentation. Three variants of the model are compared using an AM defect dataset, and the highest mean IOU achieved is 0.884, which is a good accuracy considering the various challenges in segmenting small defects. The proposed method is able to automatically segment defects in AM samples with a reliable amount of accuracy, and can be of assistance to quality control for the additive manufacturing process.

Future work could focus on tuning the network to handle the misclassification of areas with very few voxels. One improvement might involve using focal loss, which puts less penalty on well classified samples and focuses on misclassified samples (Lin et al., 2017). Dilation could be used to pre-process the training images to add voxels to the boundaries of the defects (Jackway and Deriche, 1996). The proposed models could also be trained to segment defects in AM specimens with different geometries, materials and additive manufacturing approaches. Furthermore, the effect of transfer learning on the performance and training speed of the CNN models could be explored. It would also be interesting to develop CNN models to focus on studying other important characteristics of AM defects, such as defect pattern classification.

The ability to segment defects with high accuracy could be beneficial in a number of situations in practice. A trained CNN model could be used to evaluate fabricated AM specimen for quality assurance. Furthermore, images of a specimen could be captured in the middle of fabrication to identify warning signs early on during the printing process, thereby saving materials from a failed printing process. However, to deploy this method in practice would require more training data and ensure that the model generalizes to other types of AM defects.

6 Acknowledgement and Disclaimer

This research is partially supported by the Smart Manufacturing Systems Design and Analysis Program at the National Institute of Standards and Technology (NIST), US Department of Commerce, Grant Number 70NANB19H097 awarded to Stanford University. The authors would like to thank Dr. Felix Kim of NIST for providing the 3-D printing data sets used for the experimental study.

Certain commercial systems are identified in this article. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

References

Bernsen, J. 1986. Dynamic thresholding of gray-level images. *8th International. Conference. on Pattern Recognition*, 1251–1255.

Buades, A.; Coll, B.; and Morel, J-M. 2011. Non-Local Means Denoising. *Image Processing On Line*, 1:208-12.

Buffiere, J.-Y.; Savelli, S.; Jouneau, P. H.; Maire, E.; and Fougères, R. 2001. Experimental Study of Porosity and Its Relation To Fatigue Mechanisms Of Model Al–Si7–Mg0.3 Cast Al Alloys. *Materials Science and Engineering: A*, 316(1–2): 115–126.

Caesar, H.; Uijlings, J.; and Ferrari, V. 2016. Region-Based Semantic Segmentation with End-To-End Training. *arXiv preprint arXiv 1607.07671.*

Cicek,O.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *MICCAI*, 424–432.

Ferguson, M.; Ak, R.; Lee, Y. -T. T.; and Law, K. H. 2018. Detection and Segmentation of Manufacturing Defects with Convolutional Neural Networks and Transfer Learning. *Smart and Sustainable Manufacturing Systems* 2(1): 137-164.

Ghavami, N.; Hu, Y.; Gibson, E.; Bonmati, E.; Emberton, M.; Moore, C. M.; and Barratt, D. C. 2019. Automatic Segmentation of Prostate MRI Using Convolutional Neural Networks: Investigating the Impact of Network Architecture on the Accuracy of Volume Measurement and MRI-Ultrasound Registration. *Medical Image Analysis* 58.

Gross, B. C.; Erkal, J. L.; Lockwood, S. Y.; Chen, C.; and Spence, D. M. 2014. Evaluation of 3D Printing and Its Potential Impact on Biotechnology and the Chemical Sciences. *Analytical Chemistry* 86(7): 3240–3253.

Guo, Y.; Liu, Y.; Georgiou, T.; and Lew, M. S. 2018. A Review Of Semantic Segmentation Using Deep Neural Networks. *International Journal of Multimedia Information Retrieval*, 7*(*2): 87–93

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. *IEEE International Conference on Computer Vision*, 2980-2988.

Holzmond, O.; and Li, X. 2017. In Situ Real Time Defect Detection Of 3D Printed Parts. *Additive Manufacturing*, 17: 135–142.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167.*

Jackway, P. T.; and Deriche, M. 1996. Scale-Space Properties of the Multiscale Morphological Dilation-Erosion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(1) 38–51.

Johnson, J. W. 2018. Adapting Mask-RCNN for Automatic Nucleus Segmentation. *arXiv preprint arXiv:1805.00500.*

Kim, F.H.; Moylan, S.P.; Garboczi, E.J.; Slotwinski, J.A. 2017. Investigation of pore structure in cobalt chrome additively manufactured parts using X-ray computed tomography and three-dimensional image analysis. *Additive Manufacturing*, 17:23-38.

Kim, F.H.; Moylan, S.P.; Garboczi, E.J.; Slotwinski, J.A. 2019. High-Resolution X-ray computed tomography (XCT) image data set of additively manufactured cobalt chrome samples produced with varying laser powder bed fusion processing parameters, CoCr AM XCT data, National Institute of Standards and Technology, https://doi.org/10.18434/M32162.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Lee, K.; Zung, J.; Li, P.; Jain, V.; and Seung, H. S. 2017. Superhuman Accuracy on the SNEMI3D Connectomics Challenge. *arXiv preprint arXiv:1706.00120, 2017.*

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017. Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Masad, E.; Jandhyala, V. K.; Dasgupta, N.; Somadevan, N.; and Shashidhar, N. 2002. Characterization of Air Void Distribution in Asphalt Mixes using X-ray Computed Tomography. *Journal of Materials in Civil Engineering*, 14(2), 122-129.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *IEEE International Conference on 3DVision,* 565-571.

Ngo, T. D.; Kashani, A.; Imbalzano, G.; Nguyen, K. T. Q.; and Hui, D. 2018. Additive manufacturing 3D printing: A review of materials, methods, applications and challenges. *Composites Part B: Engineering*, 143: 172–196.

Rao, P. K.; Liu, J.; Roberson, D.; Kong, Z.; and Williams, C. B. 2015. Online Real-Time Quality Monitoring in Additive Manufacturing Processes Using Heterogeneous Sensors. *Journal of Manufacturing Science and Engineering*, *137*(6): 061007 (12 pages).

Reese, R.; Bheda, H.; and Mondesir, W. 2016. Method to Monitor Additive Manufacturing Process for Detection and In-Stu Correction of Defects. *Pub . No .: US 2016 / 0271610 A1 Patent Application Publication*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI*, 234–241.

Scime, L.; and Beuth, J. 2018. A Multi-Scale Convolutional Neural Network for Autonomous Anomaly Detection and Classification in a Laser Powder Bed Fusion Additive Manufacturing Process. *Additive Manufacturing*, 24: 273-286.

Shevchik, S. A.; Kenel, C.; Leinenbach, C.; and Wasmer, K. 2018. Acoustic Emission for In Situ Quality Monitoring in Additive Manufacturing Using Spectral Convolutional Neural Networks. *Additive Manufacturing,* 21: 598–604.

Sun, W.; Brown, S. B.; and Leach R. K. 2012. An Overview of Industrial X-Ray Computed Tomography, Technical Report ENG 32, National Physical Laboratory, Teddington, Middlesex, United Kingdom.

Wolny, A. 2019. wolny/pytorch-3dunet: PyTorch implementation of 3D U-Net (Version v1.0.0). *Zenodo*. http://doi.org/10.5281/zenodo.2671581

Wu, Y.; and He, K. 2018. Group Normalization. *arXiv preprint arXiv:1803.08494*

Yu, L.; Yang, X.; Chen, H.; Qin, J.; and Heng, P. 2017. Volumetric ConvNets with Mixed Residual Connections for Automated Prostate Segmentation from 3D MR Images. *AAAI Conference on Artificial Intelligence*, 66-72.

Yu, Q.; Xia, Y.; Xie, L.; Fishman, E. K.; and Yuille, A. L. 2019. Thickened 2D Networks for 3D Medical Image Segmentation. *arXiv preprint arXiv 1904.01150.*

Zhang, B.; Liu, S.; and Shin, Y. C. 2019. In-Process monitoring of porosity during laser additive manufacturing process. *Additive Manufacturing*, 28: 497–505.

Zhou, X.; Yamada, K.; Kojima, T.; Takayama, R.; Wang, S.; Zhou, X.; Hara, T.; Fujita, H. 2018. Performance Evaluation of 2D and 3D Deep Learning Approaches for Automatic Segmentation of Multiple Organs on CT Images. *Medical Imaging: Computer-Aided Diagnosis*, 10575: 105752C.

ASSET CONDITION MANAGEMENT: A FRAMEWORK FOR SMART, HEALTH-READY MANUFACTURING SYSTEMS

Nancy Diaz-Elsayed[1]
University of South Florida,
Tampa, FL

Luis Hernandez
Global Strategic Solutions,
Vienna, VA

Ravi Rajamani
drR2 consulting, LLC,
West Hartford, CT

Brian A. Weiss
National Institute of Standards and Technology,
Gaithersburg, MD

ABSTRACT

Unscheduled downtime in manufacturing systems can be a major source of lost productivity, profits, and, ultimately, reduced process quality and reliability. However, the incorporation of asset condition management (ACM) into manufacturing systems offers an approach to improve equipment and plant operations by providing real-time condition awareness, system diagnostics, and estimates of future health to enable predictive maintenance. ACM is a framework for assessing the current and future state of health of a manufacturing system and integrating that knowledge with enterprise applications to meet the demand of production operations. In manufacturing systems, successful operations rely on the ability to maintain production assets at their optimal working levels to optimize operations and system performance. Some large corporations have made great strides in incorporating smart technologies to enhance their asset management strategy; however, small- and medium-sized enterprises (SMEs) face distinct challenges. One of the key challenges is that most SMEs do not have the wherewithal to invest in new machines nor is there standard guidance on how older machines can be integrated into an ACM solution, so that their end-to-end manufacturing process can be optimized from a health management point of view. This research presents a framework for ACM to facilitate its introduction into manufacturing systems based on their "health-ready" capabilities. Specifically, an ACM system architecture is defined for manufacturing systems, the health-ready principles and capability levels from the aerospace and automotive industries are adapted to the manufacturing domain, and the results from outreach efforts to the manufacturing community are discussed.

Keywords: Health-Ready Capability Levels, Health State; Smart Manufacturing; Diagnostics; Prognostics; Predictive Maintenance

NOMENCLATURE

ACM	Asset Condition Management
CBM	Condition-Based Maintenance
HRCS	Health-Ready Components and Systems
ICD	Interface Control Document
IIoT	Industrial Internet-of-Things
IVHM	Integrated Vehicle Health Management
PdM	Predictive Maintenance
PHM	Prognostics and Health Management
O&M	Operation and Maintenance
OEE	Overall Equipment Effectiveness
OEM	Original Equipment Manufacturer
OUC	Operational Use Case
RCM	Reliability-Centered Maintenance
RUL	Remaining Useful Life
SCADA	Supervisory Control and Data Acquisition
SME	Small- to Medium-Sized Enterprise

1. INTRODUCTION

Many machines are still reactively maintained or are operated in run-to-failure modes of operation [1]. Such an approach can result in the waste of time, cost, and resources. More advanced maintenance strategies include condition-based maintenance (CBM), reliability-centered maintenance (RCM), and the combination thereof, CBM-Plus. Continuous monitoring of an asset's condition to facilitate CBM, can help reduce unscheduled downtime and allow maintenance to be scheduled

[1] Contact author: nancyd1@usf.edu

Diaz-Elsayed, Nancy; Hernandez, Luis; Rajamani, Ravi; Weiss, Brian A. "Asset Condition Management: A Framework for Smart, Health-Ready Manufacturing Systems." Paper presented at ASME 2020 15th International Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

during a financially-opportune time with minimal impact on production operations [2]. CBM has been found to be a sustainable alternative to reactive and preventive maintenance practices [3]. At the very basic level, RCM leverages the detailed study of failure and degradation mechanisms for critical systems to appropriately schedule maintenance actions to enable the reliable operation of these systems and meet user needs. While CBM is a laudable goal, a deeper understanding of the RCM principles has done much to improve maintenance practices across many industries. One notable industry is the aerospace sector where RCM is used to determine maintenance requirements based on the analysis of the likely functional failures of components, equipment, subsystems, or systems having a significant impact on safety, operations, and life cycle cost.. Combining RCM and CBM principles, also known as CBM-Plus, is gaining traction as the U.S. Department of Defense has required its suppliers to follow this strategy [4].

In recent decades, the growth of computing power, the increased use of automation, and the advanced capabilities of sensor technology has facilitated the emergence of the Industrial Internet-of-Things (IIoT) and smart manufacturing. Readily available data has also allowed researchers to apply artificial intelligence and machine learning techniques to manufacturing applications as well. With the shift towards smart manufacturing systems, decision-makers in the factory can go beyond CBM and apply prognostics and health management (PHM) to avoid failures and sizable disturbances via predictions [2,5]. However, several challenges remain in the implementation of PHM and even more so for small- to medium-sized enterprises (SMEs), which represent the vast majority of manufacturers in the United States [6]. For example, legacy equipment likely do not have the digital capabilities that are designed into newer machines to offer plug-and-play solutions for process monitoring. Additionally, SMEs may lack the in-house expertise needed to implement and sustain smart manufacturing technologies [1] and may not have the personnel or financial resources to invest in new machines. A major challenge across manufacturers, however, is that they "lack a standard process and methodology for using [PHM] technologies on the shop floor" [7].

SAE JA6268: Design & Run-Time Information Exchange for Health-Ready Components is a standard that was recently developed to help reduce existing barriers to the successful implementation of Integrated Vehicle Health Management (IVHM) technology into the mobility sector [8]. This standard introduces the concept of "health-ready components and systems (HRCS)." HRCS, in the aerospace and automotive domains, are components that monitor and report their own health (at run-time) so that the health management of the entire aircraft or vehicle can be achieved. For the system to achieve this integrated behavior, JA6268 advocates that the supplier needs to work closely with the system integrator during the design phase to provide sufficient amount of (design-time) information to accurately assess the component's health via a higher-level "reasoning" system on the vehicle. This SAE standard has two primary objectives: (1) to encourage the introduction of a much greater degree of IVHM functionality in future vehicles at a much lower cost, and (2) to address intellectual property concerns by providing recommended design-time and run-time data specification and information exchange alternatives to help unlock the potential of IVHM [8].

This paper presents a framework to facilitate the characterization of an Asset Condition Management (ACM) system based on the "health-ready" capability of the manufacturing assets. ACM has been defined as "*the unified capability of a manufacturing system (i.e., the asset) to assess its current and future state of health and integrate that knowledge of the system state of health with enterprise applications to meet production operations demand*" [9]. This unified capability takes the form of a framework whose goal is to help manufacturers introduce IIoT technologies and advanced maintenance practices into their operations to improve the awareness of a system's health state, reduce asset downtime, and improve productivity and product quality. In this way, operation and maintenance (O&M) can be optimized and the useful life of the system can be extended. The objectives of this research were as follows: 1) define an ACM system architecture for manufacturing systems, 2) adapt the SAE JA6268 principles and Capability Level definitions to the manufacturing industry, and 3) reach out to the manufacturing community to obtain feedback on the key artifacts developed for ACM.

2. ASSET CONDITION MANAGEMENT

The first objective of this research was to define a PHM architecture reference for manufacturing systems. This research effort referenced prior work done in the aerospace and automotive industries [8,10,11] and PHM research for the manufacturing industry [2,7,12]. During this research, the use of "Asset Condition Management" was proposed for the manufacturing industry, since asset management can be recognized and interpreted by the manufacturing community without prior knowledge of PHM. Moreover, the concept behind ACM was introduced to attendees of the ASME Standards Subcommittee Meeting on Advanced Monitoring, Diagnostics, and Prognostics for Manufacturing Operations, held at NIST in May 2019. Several attendees expressed their interest in ACM to better understand its potential and how it could help their manufacturing operations.

ACM has been defined as "*the unified capability of a manufacturing system (i.e., the asset) to assess its current and future state of health and integrate that knowledge of the system state of health with enterprise applications to meet production operations demand*" [9]. ACM can support an enterprise in advancing their maintenance strategies by promoting the appropriate identification of current and desired "health-ready" capabilities throughout the system and manufacturing process. In this context, *"health-ready assets" are manufacturing systems or subsystems with the capability to monitor and report their own health*.

Maintenance practices for manufacturing equipment are based on activities that attempt to extend equipment life and reduce the likelihood of equipment failure. Scheduled and other "preventive" maintenance strategies have been the norm –

achieving maintenance objectives through regular equipment inspections and scheduled maintenance at pre-determined intervals based on operational time, cycles, units, etc. [1]. ACM aims at moving maintenance more towards Condition-Based Maintenance or Predictive Maintenance (PdM) strategies, where it makes sense. In some assets, the preventive and run-to-failure strategies may be enough to meet the enterprise business objectives. CBM and PdM maintenance strategies require the monitoring and management of the condition of the manufacturing equipment to avoid disruptions in operations due to equipment downtime. The objective of ACM is to drive maintenance when it is needed to ensure safety, reliability, availability, and reduced life cycle costs. In practice, these strategies help reduce unnecessary downtime by employing "just-in-time" maintenance procedures.

2.1 A Physical Hierarchy for Manufacturing Systems

To decompose the complexity of manufacturing systems, a physical hierarchy is defined ranging from an enterprise to a component level as represented by Figure 1. Weiss et al. [13] previously described the functions of the entities at each of the levels in terms of activity diagrams, and identified the monitoring and PHM functions at each level. The operational metrics related to PHM at each level flow up to be aggregated at the top level to get an overall picture of the health of the entire system.

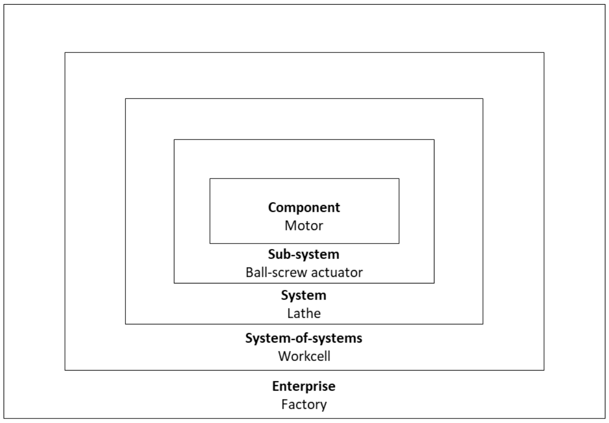


FIGURE 1: THE PHYSICAL HIERARCHY OF MANUFACTURING SYSTEMS

Examples of a physical entity are provided in Figure 1 for each category. At the highest level is the enterprise, which can represent a factory or the larger environment in which the factory operates, such as a collection of factories connected by logistic chains. Next is a workcell or a production line (a system-of-systems). In this case, the "System" is represented by a lathe, and the downstream "sub-system" is a linear actuator. The actuator consists of several components such as the ball-screw, the motor, the motor controller, the tool assembly, etc. Components can be further broken down into sub-components, but for illustrative purposes, the "Enterprise" to "Component" levels are shown in Fig. 1. Also, as has been pointed out by Weiss et al. [13], this is a relative hierarchy. If the focus is at a different level, a component might well become a system, such as what might happen for a motor original equipment manufacturer (OEM).

To understand how one key element of ACM may work in this context, consider the function of estimating the health of the ball-screw actuator. Let us assume that measuring the slew-rate of the actuator is enough to adequately assess one aspect of the actuator health (e.g., the time it takes for the actuator to slew from one position limit to the other and back is enough to determine the health of the system). As all components age, the slew rate increases for the same motor command inputs. By trending this rate, one can calculate a metric indicative of the health of the actuator.

In this example, it is assumed that the slew rate is measured in the computer by measuring the time it takes for the actuator to travel between limit switches. This means that the signals from the switches must be acquired, sampled and encoded, then transmitted to a computer where the analysis is done. Some physical elements are very clear in this implementation. The limit switches are key sensing elements, as are their companion data acquisition electronics, which includes the A/D convertors, for example. In this embodiment, it is assumed that digital data is further transmitted to another central computer for analysis. It is quite conceivable that in another embodiment, the analysis is done in the same location that data is acquired. Such distributed architectures will become more common as distributed computational systems become more feasible with advanced technology, such as higher throughput and faster communications. We present this example, because a lot of interest has been generated in recent decades on the health management of electromechanical systems such as the linear actuator in the mobility sector. This is mainly because of the increase in electrification both of automobiles and aircraft. This trend will continue, and the manufacturing sector can make full use of it to increase its ACM capabilities. To situate this example in the ACM context we compare this with what would be done traditionally in monitoring an actuator of this kind, which is next to nothing! Any estimate of the health of the actuator would be inferred from other parameters further downstream. In a highly ACM-capable manufacturing system, the parameters from the actuator itself will be monitored and its health directly measured. This would be fed to a higher system health management function that would use this lower level intelligence to assess the health of the entire system. The more integrated these capabilities are, the higher the level of health management capability we would assign the system.

2.2 ACM System Architecture Reference

Leveraging sensors and "health-ready systems" in the ACM framework calls for practitioners to focus on identifying and defining information that must be exchanged between the levels in the physical hierarchy or between lower-level components, sub-systems and the higher-level system functions. Hence, data transfer will occur between three tiers: a) Production Assets, Sensors, and Health-Ready Systems, b) Data Acquisition, Contextualization, Integration and Exchange, and c) Enterprise

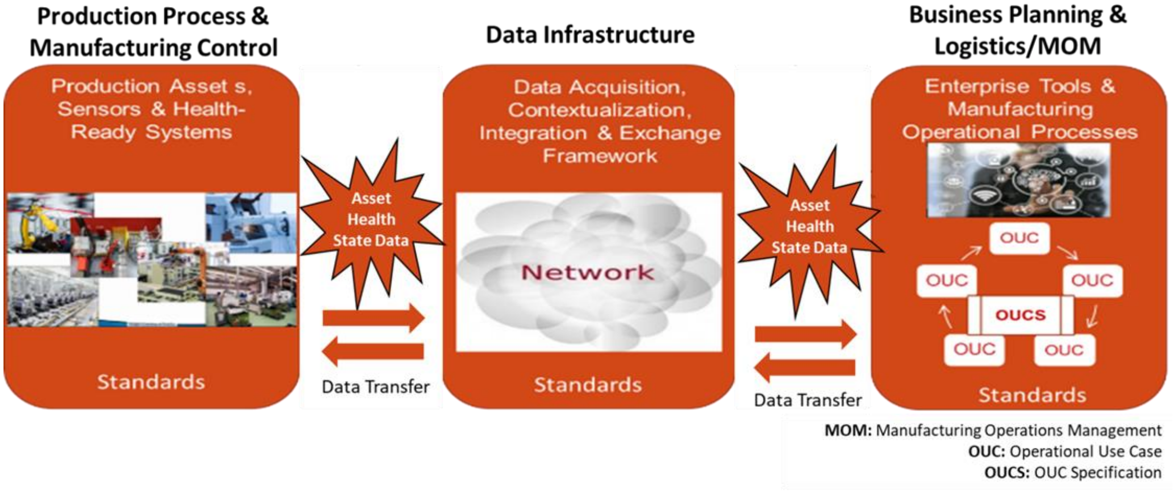

Production Process & Manufacturing Control
Production Asset s, Sensors & Health-Ready Systems
Standards
Asset Health State Data
Data Transfer
Data Infrastructure
Data Acquisition, Contextualization, Integration & Exchange Framework
Network
Standards
Asset Health State Data
Data Transfer
Business Planning & Logistics/MOM
Enterprise Tools & Manufacturing Operational Processes
OUC
OUC
OUC
OUCS
OUC
OUC
Standards
MOM: Manufacturing Operations Management
OUC: Operational Use Case
OUCS: OUC Specification

FIGURE 2: AN OVERVIEW OF THE ACM SYSTEM ARCHITECTURE

Tools and Manufacturing Operational Processes (see Figure 2). The ACM system functions can be integrated with the ANSI/ISA-95 Level 0-1-2 functions (production process and manufacturing control) to enable the automation of equipment health data. Knowledge of the equipment state of health may then drive efficiencies and the effectiveness of Level 3-4 (manufacturing operations management and business planning and logistics) functional operations including Maintenance Operations, Production Operations, Quality Operations, Plant Production Scheduling, etc. The ANSI/ISA 95 series of standards address the integration of the enterprise operations with the manufacturing control system [15]; this approach aligns quite well with the concept of ACM.

The Data Acquisition, Contextualization, Integration and Exchange functions can be local (to the specific equipment or process), distributed (e.g., across the factory), or cloud-based (i.e., across the enterprise). The operational use cases (OUCs) express the specific business purpose for using the asset health state data captured by the system; they are used to convey user-specific needs and drive ACM system functional requirements. For example, from an operational perspective, the specific value-added a maintenance department expects from the capability for assessing the individual equipment health, the health of subsystems or workcells.

The objective of this “top-down” approach is to associate the value of ACM with overall equipment effectiveness (OEE) to specify and characterize manufacturing productivity based on equipment and process health. OEE is a commonly used metric across many manufacturing facilities that indicates performance based upon measured productivity, quality, and availability metrics and measures [14]. OEE can be derived at multiple levels of the manufacturing hierarchy, from the enterprise level down to the equipment level. The OUC specification also drives the communication between the business and ACM development teams and provides a way to represent user requirements that align with the operational business requirements. They should be specified in non-technical language and enable ordering, grouping, and prioritization. Also, the use case specification is the primary input for user acceptance testing as well as the development of test cases. Nevertheless, the specific architecture solution depends on the motivating factors driving the need for ACM, the existing infrastructure, and the economics of the situation.

Note that a system implementation may utilize all or a portion of the health-ready functions inherent in the production assets. The reason is that some sensor parameters and lower level diagnostic codes are representative of symptoms and are manifested in higher-level health-state indications. Specifying how each layer in the system produces or uses the data available for exchange was outside of the scope of this research, but future work should leverage existing standards and recommend how data capture, processing, and exchange within and between the architecture tiers should occur for ACM in greater detail.

3. HEALTH-READY CHARACTERIZATION

Historically, manufacturing equipment maintenance policies have been defined based on maintenance actions that attempt to extend equipment life and minimize likelihoods of equipment failure, e.g., preventive maintenance strategies relying on scheduled tasks. ACM can support an enterprise in advancing its maintenance strategies by promoting the appropriate identification of current and desired “health-ready” capabilities throughout the system and manufacturing process.

Starting with data capture, an ACM practitioner characterizes each process or function using technical specifications and interface control documents (ICD). The specific functions and processes depend on the design and implementation of the ACM capabilities considering the physical scope (e.g., the system, sub-system or component level) and the type of equipment involved (e.g., new assets with advanced monitoring capabilities or legacy equipment without them). For example, a certain component might have perfect data acquisition capability, i.e., it can reliably deliver the appropriate data to an acquisition system for a health assessment. This data capture system would buffer the data, convert it as needed and deliver it to the appropriate receivers for further processing (e.g., by a downstream SCADA system). The assessment might end at this point, or, for a more complex sub-system, data processing might be included in the health assessment as well. Such a sub-system might have built-in data normalization and parameter correction, and have the capability of providing such normalized data to the more advanced ACM functions further downstream. All diagnostic and prognostic analysis systems need to work with processed data. Accordingly, most PHM systems have built-in normalization routines to filter noise and correct for different operating conditions. This way, data under different operating conditions can be compared. In the aerospace domain, for example, this would include correction for standard day conditions, and varying loading factors. The functions or processes involving state detection, health assessment and prognostics assessment might not be very advanced. In fact, it is expected that most assets will not have prognostics capabilities (i.e., capabilities to predict future health states). But for some smarter components, there may be limited ability to assess the current state of health, especially in terms of detecting anomalies and faults. With historical data and analytics, the ability to extrapolate the estimates and to assess remaining useful life (RUL) may also be provided, which would generate the data needed for predictive maintenance. Future efforts will involve some measures of validation.

3.1 Health-Ready Capability Levels

Adapting the SAE JA6268 IVHM Capability definitions [8], a proposed approach for establishing ACM Capability Levels for smart manufacturing systems was developed. The definitions consist of a progression of ACM Capability Levels (from Level 0 to Level 5), which are based on functional aspects of the asset's inherent capabilities:

- **Capability Level 0 - Limited Failure Indicators:** Asset maintenance is prompted by either scheduled preventive maintenance or inspections, when the asset operator is alerted by failure indicator lights or gauges conveying limited awareness, or when the operator observes a performance issue.

- **Capability Level 1 - Diagnostics**: Asset is equipped with diagnostic functions. Maintenance personnel gain diagnostic insight by viewing or extracting operating parameters and/or diagnostic information from the asset. Simple (e.g., relatively high-level) fault isolation information is available.

- **Capability Level 2 - Asset Monitoring**: Asset is equipped with the ability to automatically capture data, possibly store the data, and diagnose based on intelligent algorithms. A key characteristic is that data can be used to monitor real-time performance or to capture performance history over time for subsequent analysis. More detailed fault isolation information is available.

- **Capability Level 3 - Prognostics**: Asset operator and maintenance personnel are provided with alerts of impending faults, listing severity levels, along with estimated RUL, and recommended fault remediation and maintenance actions.

- **Capability Level 4 - Comprehensive ACM**: Asset operator and maintenance personnel are provided with diagnostics and prognostics information at the enterprise level with alerts of impending faults listing severity levels, RULs, and recommended fault remediation and maintenance actions. Limited logistics recommendations may also be provided. This would include resource allocation, resource scheduling, and spare part locations, which may be used to support transportation decisions for tools and material.

- **Capability Level 5 - Self-Adaptive ACM**: ACM capability is integrated with asset control and enterprise management to automate logistics and maintenance scheduling based on available information, resources, and costs. Assets can be automatically (or manually, based on automated advice) reconfigured and repurposed to deliver acceptable performance in the presence of asset or process degradation, with detailed advice on fault remediation and system maintenance.

A key transition occurs between Level 2 and Level 3, where prognostics and predictive analytics are brought to bear to significantly enhance the ACM capability. Moreover, as the levels increase, the maintenance practices shift. For example, maintenance at Level 0 would occur after an inspection or failure. At Level 1 and 2, scheduled maintenance would be coupled with CBM (driven by diagnostic functions) but limited predictive capabilities would be available. At Levels 3 and 4, maintenance would rely on predictions of the asset's condition (i.e., predictive or CBM). Lastly, Level 5 applies CBM with logistics optimization and adaptive process control. Referring to the example of the ball-screw actuator, typical systems are at Level 0 with little or no diagnostic capability. The monitoring of the slew rate as described above would constitute a basic information source that could be used to develop capability at Level 1 or 2 depending on how the information is used. With more sophisticated algorithm that trends the slew rate over time and extrapolates it to estimate RUL, we may even be able to get Level 3 capability. To get to higher levels, the system-wide monitoring and assessment capability must be in place so that the information from the actuator can be used for more than just prediction but also automated maintenance scheduling, etc.

3.2 The Health-Ready Capability Levels in Practice

The outreach to manufacturers was conducted via a survey that was developed and deployed in Qualtrics, as well as one-on-one interviews. The guidelines presented in Sections 2 and 3 incorporate the changes made following the feedback from the outreach efforts. In all, feedback was obtained from six manufacturing entities including four SMEs, one OEM, and one manufacturing trade association.

After being introduced to the ACM definition and the health-ready capability levels (including details about the data acquired and the maintenance practices at each capability level), manufacturing entities were asked which health-ready capability level was achieved by most of their assets and the highest level achieved by their assets. While most assets operated at a Level 2, one manufacturer did operate equipment that achieved a Level 5 as shown in Figure 3. The asset that achieved a Level 5 was a CNC machine tool that had advanced monitoring and control capabilities used for precision manufacturing operations. The machine could self-calibrate based on real-time conditions and optimize the life of the ball-screw, bearings, and rails. In the case where "I don't know" was selected, the manufacturer was an OEM and was unaware of the highest capability level achieved by their suppliers.

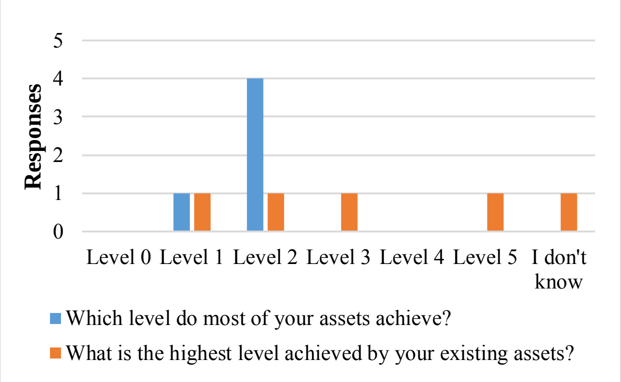


FIGURE 3: ACM CAPABILITY LEVELS OF EXISTING ASSETS; SUMMARY OF RESPONSES FROM THE MANUFACTURERS (EXCLUDES THE TRADE ASSOCIATION).

To further investigate how these assets were maintained, the manufacturers were also asked who the individual or organization was that maintained the assets at the highest level. Table 1 shows the results for all surveyed manufacturers. While routine maintenance was often performed by internal personnel, the supplier of the asset, component, or technology was also a key contributor to maintenance activities as summarized in Table 1. The latter played a key role in maintaining the asset at Level 5 where more sophisticated tasks and deeper knowledge of the asset were necessary.

TABLE 1: THE INDIVIDUAL OR ORGANIZATION THAT MAINTAINED THE ASSET(S) WITH THE HIGHEST ACM CAPABILITY LEVEL AT THEIR FACILITY.

Health-Ready Capability Level	Individual or Organization Maintaining the Asset(s)
Level 0	Not Applicable
Level 1	The supplier of the asset, component, or technology
Level 2	Internal personnel (e.g., technician)
Level 3	The supplier of the asset, component, or technology
Level 4	Not Applicable
Level 5	Internal personnel for routine maintenance and the supplier of the asset for calibration and annual checks.

4. CONCLUSION

Key challenges for SMEs to advance their maintenance strategy include not having the wherewithal to invest in new machines nor standard guidance on how older machines can be integrated into an ACM solution, so that their end-to-end manufacturing process can be optimized from a health management point of view. Legacy equipment may not have the digital capabilities that are designed into today's smart machines, but that does not mean it is impossible to extract high-quality, meaningful data from these assets. Many SMEs may not be realizing the full benefits of ACM because they have older machines and the traditional approach assumes that there is no health-ready capability within the assets. That is, it is assumed that these (legacy) machines don't have the needed sensors to track data or the necessary capabilities to easily connect the assets to a central repository. By retrofitting these systems with sensors and data acquisition units, they may be brought into higher ACM capability levels without excessive investments; the benefits derived from enhancing the capabilities can far outweigh the costs.

The ACM framework offers an opportunity to extend the useful life of manufacturing systems, such that waste and resource consumption can be reduced. It is not the intention of this research to suggest that all manufacturers should maintain their assets at a Level 5. Instead, with the framework in hand, manufacturers can benchmark which health-ready capability level their assets achieve. By comparing the capabilities achieved by alternative systems, they can optimize their decision-making based on the factors that are most critical to their factory or enterprise; for example, the value added, the resources consumed, and the ability to meet their customer's demand.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Institute of Standards and Technology (NIST) (Cooperative Agreement No. 70NANB18H260). The authors wish to acknowledge the cooperation and support of the Prognostics and Health Management for Reliable Operations in Smart Manufacturing project team at NIST for the collaborative discussions and their valuable inputs during the performance of this work. The project team would also like to acknowledge the Center for Advanced Vehicular Systems Extension (CISE) at Mississippi State University, Crevalle Boats, Custom Manufacturing & Engineering, Inc., Custom Metal Designs, Inc., Ferrera Tooling, FloridaMakes, the MTConnect Institute, and the Tampa Bay section of the Society of Women Engineers for their support of this project during the manufacturer outreach process.

NIST DISCLAIMER

The views and opinions expressed herein do not necessarily state or reflect those of NIST. Certain commercial entities, equipment, or materials may be identified or referenced in this manuscript in order to illustrate a point or concept. Such identification or reference is not intended to imply recommendation or endorsement by NIST; nor does it imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] Helu, M., and Weiss, B., 2016, "The Current State of Sensing, Health Management, and Control for Small-to-Medium-Sized Manufacturers," *ASME 2016 11th International Manufacturing Science and Engineering Conference, MSEC 2016*, American Society of Mechanical Engineers.

[2] Vogl, G. W., Weiss, B. A., and Helu, M., 2019, "A Review of Diagnostic and Prognostic Capabilities and Best Practices for Manufacturing," J. Intell. Manuf., **30**(1), pp. 79–95.

[3] Nezami, F. G., and Yildirim, M. B., 2013, "A Sustainability Approach for Selecting Maintenance Strategy," Int. J. Sustain. Eng., **6**(4), pp. 332–343.

[4] U.S. Department of Defense, 2012, *Department of Defense Instruction: Condition Based Maintenance Plus (CBM+) for Materiel Maintenance*.

[5] Lee, J., Ghaffari, M., and Elmeligy, S., 2011, "Self-Maintenance and Engineering Immune Systems: Towards Smarter Machines and Manufacturing Systems," Annu. Rev. Control, **35**(1), pp. 111–122.

[6] Manufacturing Institute, 2019, "Small Companies Dominate the Industrial Landscape" [Online]. Available: http://www.themanufacturinginstitute.org/Research/Facts-About-Manufacturing/Economy-and-Jobs/Company-Size/Company-Size.aspx. [Accessed: 18-Oct-2019].

[7] Jin, X., Weiss, B. A., Siegel, D., and Lee, J., 2016, "Present Status and Future Growth of Advanced Maintenance Technology and Strategy in US Manufacturing," Int. J. Progn. Heal. Manag., **7**(Spec Iss on Smart Manufacturing PHM).

[8] SAE International, 2018, *JA6268: Design & Run-Time Information Exchange for Health-Ready Components*.

[9] Hernandez, L., Diaz-Elsayed, N., and Rajamani, R., 2019, *Final Performance Report: Characterization of PHM Technologies for Manufacturing at Small and Medium Sized Enterprises*.

[10] SAE International, 2016, *ARP6803: IVHM Concepts, Technology and Implementation Overview*.

[11] SAE International, 2012, *ARP6290: Guidelines for the Development of Architectures for Integrated Vehicle Health Management Systems*.

[12] Jin, X., Siegel, D., Weiss, B. A., Gamel, E., Wang, W., Lee, J., and Ni, J., 2016, "The Present Status and Future Growth of Maintenance in US Manufacturing: Results from a Pilot Survey," Manuf. Rev., **3**.

[13] Weiss, B. A., Sharp, M., and Klinger, A., 2018, "Developing a Hierarchical Decomposition Methodology to Increase Manufacturing Process and Equipment Health Awareness," J. Manuf. Syst., **48**, pp. 96–107.

[14] Muchiri, P., and Pintelon, L., 2008, "Performance Measurement Using Overall Equipment Effectiveness (OEE): Literature Review and Practical Application Discussion," Int. J. Prod. Res., **46**(13), pp. 3517–3535.

[15] International Society of Automation, 2018, *ISA95, Enterprise-Control System Integration*.

ASSESSMENT OF A NOVEL POSITION VERIFICATION SENSOR TO IDENTIFY AND ISOLATE ROBOT WORKCELL HEALTH DEGRADATION

Brian A. Weiss[1], Jared Kaplan
National Institute of Standards and Technology (NIST)
Gaithersburg, Maryland, USA

ABSTRACT

Manufacturing processes have become increasingly sophisticated leading to greater usage of robotics. Sustaining successful manufacturing robotic operations requires a strategic maintenance program. Without careful planning, maintenance can be very costly. To reduce maintenance costs, manufacturers are exploring how they can assess the health of their robot workcell operations to enhance their maintenance strategies. Effective health assessment relies upon capturing appropriate data and generating intelligence from the workcell. Multiple data streams relevant to a robot workcell may be available including robot controller data, a supervisory programmable logic controller data, maintenance logs, process and part quality data, and equipment and process fault and failure data. This data can be extremely informative, yet the sheer volume and complexity of this data can be overwhelming, confusing, and sometimes paralyzing. Researchers at the National Institute of Standards and Technology have developed a test method and companion sensor to assess the health of robot workcells which will yield an additional and unique data stream. The intent is that this data stream can either serve as a surrogate for larger data volumes to reduce the data collection and analysis burden on the manufacturer; or add more intelligence to assessing robot workcell health. This article presents the most recent effort focused on verifying the companion sensor. Results of the verification test process are discussed along with preliminary results of the sensor's performance during verification testing.

Keywords: degradation, diagnostics, industrial robot systems, kinematics, manufacturing, prognostics, prognostics and health management (PHM), use cases, verification, workcell.

NOMENCLATURE

6DOF	Six-Degree-of-Freedom
COTS	Commercial-off-the-shelf
NIST	National Institute of Standards and Technology
PHM	Prognostics and Health Management
PLC	Programmable Logic Controller
PdM	Predictive Maintenance
PM	Preventive Maintenance
PVS	Position Verification Sensor with Discrete Output
RM	Reactive Maintenance
SS	Stainless Steel
V&V	Verification and Validation

1. INTRODUCTION

Every product used in daily life is manufactured in some way; from the clothes that are worn, to the automobiles that are driven, and the cell phones that are used. Each of these products, and many more, are being offered with more customizable configurations and options. Coupling the concept of product customization with the evolution of manufacturing technologies has led to manufacturing processes becoming increasingly flexible and complex [1-3]. To enable greater flexibility and handle increased task complexity, manufacturers have turned to robotic technologies. Robots can offer manufacturers numerous benefits (e.g. increase accuracy, precision, repeatability, efficiency) as compared to conventional manufacturing automation or manual capabilities [4-6].

As robots have become more adept at providing a broader range of manufacturing capabilities, they are presenting more operational complexity. More complexity typically leads to greater opportunity for faults and failures in the process and equipment. In turn, this spurs more unplanned maintenance or more planned maintenance routines to lessen the potential for unexpected faults or failures [7-9]. Manufacturers, from small to large, recognize that strategic and scripted maintenance

[1] Contact author: brian.weiss@nist.gov

Weiss, Brian A.; Kaplan, Jared. "Assessment of a Novel Position Verification Sensor to Identify and Isolate Robot Workcell Health Degradation." Paper presented at ASME 2020 15th International Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

strategies are critical to maximize process and equipment availability [10, 11].

Personnel from the National Institute of Standards and Technology (NIST) are conducting research to develop measurement science products (e.g., test methods, performance metrics, reference datasets) to promote the verification and validation of monitoring, diagnostic, and prognostic technologies within manufacturing operations as part of NIST's Smart Manufacturing programs [12-14]. These measurement science products are intended to decrease equipment and process downtime and increase reliability for smart manufacturing systems. A critical output of the effort is the dissemination and adoption of these products by the manufacturing community. A specific measurement science product being developed is a test method, along with a companion sensor to be used within the test method, to assess the degradation of the workcell's kinematic chain (i.e. the assembly of individual rigid bodies at joints) [15, 16]. When applied by industry to operational manufacturing robot workcells, the test method can identify and isolate degradations, within the workcell, that are negatively impacting the accuracy of the process. Ultimately, the test method, along with the sensor, will be made publicly-available to industry for widespread adoption and implementation. Before this can be achieved, the overall test method, including the sensor, require testing and verification.

This publication presents NIST's latest efforts to verify the performance of the companion sensor. Early work focused on manually testing the sensor. This proved overly time-consuming and prone to human error. Current efforts have focused on automating the testing process to cover more test points in a faster manner, as compared to manual testing. The remainder of this article is organized as follows. The Background section presents the motivation for this effort. The Test Method and Sensor Description section describe the test method in detail and discuss how the sensor is integrated with the test method. The Experimental Design section talks about the design and development of the automated measurement test stand. The Testing and Results section highlight the tests that were conducted. The Lessons Learned Section analyzes the preliminary results from two perspectives – how well did the automated measurement test stand perform and how well did the companion sensor perform. Lastly, the Future Efforts section describes the expected next steps of this research effort.

2. BACKGROUND

Maintenance is typically a critical factor to ensure long-term, acceptable operation for any process or piece of equipment. Although important, maintenance can at times be expensive and cumbersome [17, 18]. Without sufficient intelligence regarding the health of their manufacturing operation, which would include resident robot workcells (defined to include the robot, end-of-arm tooling, sensors, controller(s), and other supporting automation), manufacturers cannot make informed decisions about the most cost-effective and efficient maintenance strategy(ies) by which to maintain their operations. The discipline of monitoring, diagnostics, and prognostics to enhance decision-making surrounding maintenance activities is known as prognostics and health management (PHM) [19-22]. The optimal maintenance schedule includes striking a balance between preventative maintenance (PM) (performing specific maintenance activities at set intervals of time, cycles, or other measurable unit) and predictive maintenance (PdM) (performing specific maintenance activities only when the current condition of equipment predicts that maintenance is necessary) all in an effort to minimize reactive maintenance (RM) ("fix it when it breaks") [23]. Preventive maintenance can become excessive leading to wasted time and money. However, minimizing or trivializing preventive maintenance can lead to too much reactive maintenance. Unexpected shutdowns resulting from reactive maintenance can lead to lost revenue and even lost customers.

Tracking the health of robot workcells allows manufacturers to make informed decision about the type of maintenance needed and the schedule at which it should be performed. Data is required to generate intelligence regarding robot workcell health. Multiple data streams are likely to exist that are relevant to identifying the health of a robot workcell. They include:

- Robot-level data – Many robot controllers capture a range of information including robot joint and tool center positions, velocities, and accelerations; joint temperatures; joint currents; and joint voltages. This data is usually quantitative.
- Process-level data – This type of data could be captured from the overall process controller or from a supervisory Programmable Logic Controller (PLC). Time data is an example of process-level data, where time can include overall task time, sub-task time, and takt time. This data is usually quantitative.
- Quality data – As measured of the part produced once the workcell has completed its operation. This data could be quantitative or qualitative.
- Operational configuration data – Information describing the workcell's configuration and operations. This can include the make and model of equipment, technologies, and sensors critical to the workcell's function. This information can also describe the workcell's operation (e.g., *Robot Arm 01 lifts a 2 kg box, puts it down, then lifts a 5 kg box, puts it down, etc.*). This can be both quantitative and qualitative descriptive data.
- Fault and failure data – For any fault or failure that presents itself within a manufacturing operation, it is almost always documented. This documentation could be done by multiple individuals including the equipment and process operator, maintenance technician, or supervisor. Fault and failure data can be quantitative from sensor and equipment readings (e.g., the current peaked at 18 A). It can also be descriptive – e.g., smoke started rising from the motor, I then heard a loud crack, and the motor stopped running.

- Maintenance logs – Most manufacturers document their maintenance activities, whether it results from planned maintenance (predictive or preventive) or from unplanned maintenance (reactive). Maintenance records can include what specific work was performed and a descriptive of the restorative state that the equipment is now in.

Effectively turning all of this data into health intelligence is non-trivial, especially as it relates to robot workcells [24]. Multiple algorithms exist that dictate how to fuse or analyze the data to discover new intelligence. Each method comes with its advantages and disadvantages. Generating the appropriate amount of intelligence can be paramount to an organization. Too little intelligence can lead to vast uncertainty and present an excessive amount of options to maintenance decision-makers. An over-abundance of intelligence can be costly, both in the time to generate and required resources. For highly critical processes or equipment, a cost-benefit analysis could dictate the need for more, as opposed to less, intelligence. The right amount of intelligence can prevent degradations that force part quality or process productivity to unacceptable levels while still promoting cost effective operations [25, 26]. NIST has developed the *Identification and Isolation of Robot Workcell Accuracy Degradation* test method and companion Position Verification Sensor with Discrete Output (U.S. Patent Pending, Serial Number 16/572,847) to provide a new stream of direct intelligence that can localize where faults and failures are occurring in a workcell's kinematic chain; and offer additional data that can be fused with the afore-mentioned data to offer a richer understanding of the workcell's health. The following section describes the test method, position verification sensor, and the need for sensor verification.

3. TEST METHOD AND SENSOR DESCRIPTION

The *Identification and Isolation of Robot Workcell Accuracy Degradation* test method was developed in concert with a multi-robot testbed at NIST [27]. FIGURE 1 presents the NIST's PHM for Robot Systems testbed with many representative features of a manufacturing robot workcell including robots, controllers, and end-of-arm tools. The testbed features two six-degree-of-freedom (6DOF) Universal Robots, each with industrial controllers. The smaller robot, a UR3 (shown on the left in FIGURE 1), is configured to perform path planning operations including drawing several unique patterns) while the larger robot, a UR5 (shown on the right in FIGURE 1), is configured to perform a material handling operation. The UR3 is configured with a pen at its tool-center-position using a specially-designed mount. The UR5 is configured with an RG2 gripper that is controlled through the UR5's controller. Supervisory commands are issued to both the UR3 and UR5 by a single Beckhoff PLC.

The testbed's primary manufacturing-relevant use case is for the larger robot to physically place test parts on fixtures within reach of the path planning robot for this smaller robot to draw on the test part. This drawing activity is analogous to welding, adhesive application, and additive manufacturing processes performed by six-degree-of-freedom robot systems. Additional information on the testbed and use case is shown in [15, 27]

The *Identification and Isolation of Robot Workcell Accuracy Degradation* test method involves testing the positioning repeatability, and therefore the health, of different components along the robot workcell's kinematic chain. The test method uses the NIST-developed Position Verification Sensor with Discrete Output (PVS) sensor (depicted in FIGURE 2) to measure the positioning of the components in question [16]. The test method is robot agnostic; it can be integrated and executed with any 6DOF robot workcell. When different measurement points along

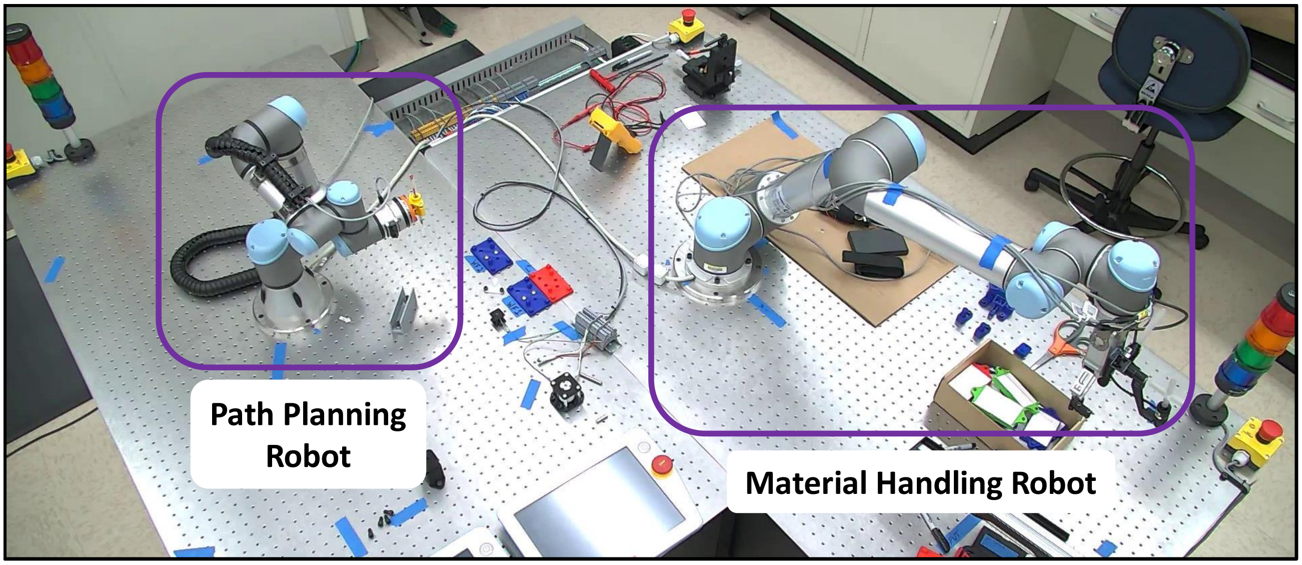


FIGURE 1: NIST PHM FOR ROBOT SYSTEMS TESTBED

Weiss, Brian A.; Kaplan, Jared. "Assessment of a Novel Position Verification Sensor to Identify and Isolate Robot Workcell Health Degradation." Paper presented at ASME 2020 15th International Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

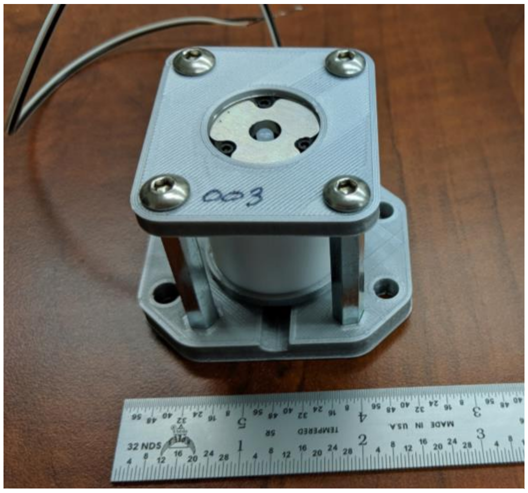


FIGURE 2: POSITION VERIFICATION SENSOR WITH DISCRETE OUTPUT

the kinematic chain are tested, the test method's results aim to isolate any source(s) of the error so the responsible component can be appropriately addressed. Implementing the test method first involves affixing mechanical keys at strategic locations along the workcell's kinematic chain that can present positional degradation. The PVS is physically installed within the workcell's work volume. The keys are commercial-off-the-shelf (COTS) items with precise geometries. The dimensions of the keys are in English units so the sensor's clearance dimension is manufactured in English units.

In the case of the NIST testbed, keys have been attached to multiple elements including the robot's tool flange, physical mount of the gripper jaws, and the movable gripper jaws, themselves (shown in FIGURE **2** 3). Once the keys are installed, the robot, along with the other movable elements in the kinematic chain (e.g., gripper), are systematically programmed to 'insert' their keys into the PVS. The PVS is designed such that the key can be successfully inserted within a specific tolerance as determined by the dimensions of the key and the opening of the PVS. If the key's position is inaccurate beyond the tolerance of the PVS' opening, then the test of that specific point will fail. Otherwise, if the movement of the key is within the tolerance of the PVS' opening, then the key will be successfully inserted into the PVS and this specific point will pass. The element to which the key is attached when a failure is observed highlights the approximate location of the degradation within the kinematic chain. More details are available in [16, 28].

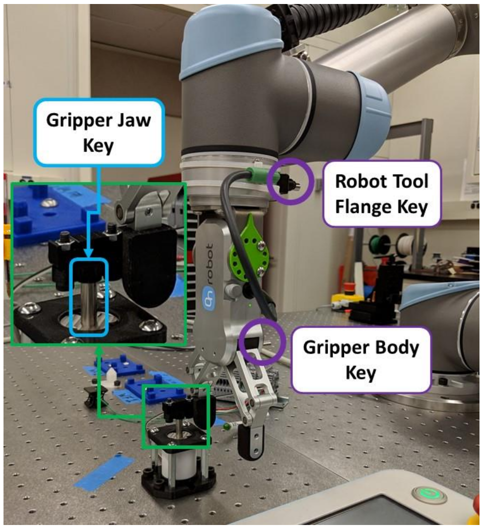


FIGURE 3: NIST TESTBED KEY MOUNTING LOCATIONS

The insertion of the key into the PVS is being verified to determine what uncertainty, if any, exists regarding the known tolerances of the fit. For example, if the clearance between the key and the PVS are designed to 25 µm, then the ideal scenario is that the key has a range of motion of 25 µm regarding its insertion into the PVS. Realistically speaking, uncertainties exist including manufacturing tolerances of the key and the PVS, the setup error of the key with respect to the PVS, etc.

As noted earlier, the PVS provides binary output, meaning that the element being tested either passes or fails its specific positioning test. If all elements can insert the key into the PVS then the workcell is considered healthy to the designed tolerance between the key and the PVS. If an element fails, then that data can be used to determine specifically where within the workcell a change or degradation has occurred. By itself, or coupled with other data from the workcell's operations, this data can be used to then efficiently respond to that change. The declaration that the workcell is healthy to the designed tolerance is only true if all uncertainties are known and quantified. If not, for example, a kinematic chain that successfully passes all tests with a tolerance of 25 µm may only be healthy to 50 µm or larger. Whatever it may be, it is critical for the manufacturer to know exactly what they are testing. The verification testing of the PVS is a core step in ultimately determining the uncertainty of the overall test method.

4. EXPERIMENTAL DESIGN

The PVS' performance needs to be verified to understand its capability and measurement uncertainties as it's used within the test method. The PVS is actively being tested to gather this information. A verification test process has been developed using an automated, linear, three-axis stage. The test stand is set up such that a standard key is mounted to the three-axis stage and a PVS is mounted directly below (as shown in FIGURE 4). The stage can then be commanded to move and insert the key into the PVS at different locations in an attempt to achieve a positive response from the PVS (i.e., the key is successfully inserted into the sensor). This test stand simulates robot movement with an attempt to insert the key into the PVS and the repeatability uncertainties that exist within them.

This test process was previously performed manually where there was a desire to automate the process because of the manual process being tedious, cumbersome, and susceptible to human

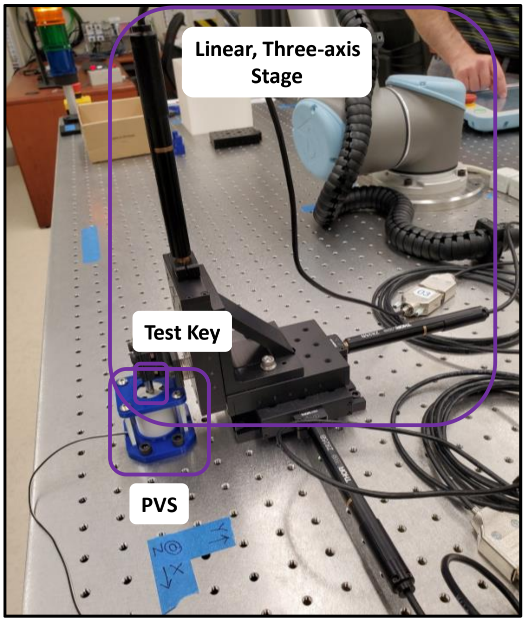


FIGURE 4: AUTOMATED TEST STAND

error. Stage movement had previously been controlled using manual micrometers, for the X and Y axis, and a lever for the Z axis (shown in FIGURE 5). The X and Y movements test different locations or points of the sensor while the Z axis movement is used to engage or disengage the key from the PVS. The chosen units enabled the user to move the stage in increments of one micrometer at a time providing very granular control of the stage's movements. To automate the movement of the stage, the micrometers and lever have been replaced with three motor controllers that are connected to three actuators (shown in FIGURE 4). This new hardware has the same range of motion and resolution as the manually-driven stage.

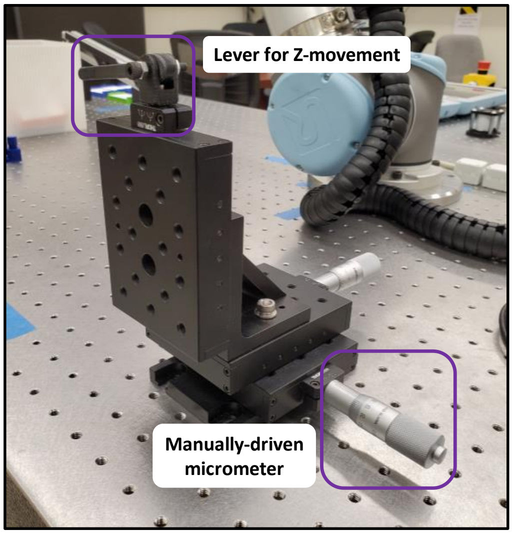


FIGURE 5: MANUAL TEST STAND

A Matlab program (running off a laptop PC that is connected via USB to the motor controllers of the linear stage) provides a test grid encompassing the top hole of the PVS. This grid consists of points to be tested, with the size and resolution of the grids (i.e., number of points to be tested and spacing between points) determined by user input. The program then selects the order for which the test points are evaluated by using one of three user-selected methods (i.e., random pattern, vertical movement testing each Y-column before moving onto the next, or horizontal movement testing each X-row before moving onto the next) and commands the stage to move the key to those points. At each point, the stage will attempt to insert the key into the PVS. The PVS' digital output is connected to an oscilloscope, which is used to determine the success of each attempted insertion. The oscilloscope measures the output voltage of the sensor at each test point, which is then input into the Matlab program. The oscilloscope is connected to the same PC laptop via USB connection. Upon receiving the signal input from the oscilloscope, the program can determine if the insertion was successful or not based on the captured voltage value (high voltage indicates success). Each test result is recorded within the program. Sample output from the Matlab is shown in FIGURE 6. Movements of the stage are programmed via Matlab in English Units given the English dimensions of the COTS key.

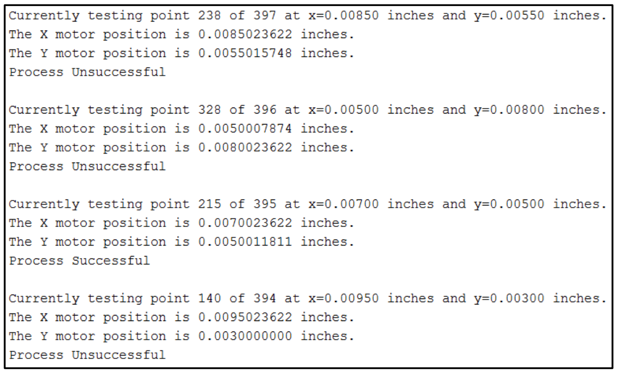
The results of the insertion at each point are stored, formatted into grids, and exported into MS Excel ®. Each point in the grid contains either a green "1" (indicating a successful test point) or a red "0" (indicated failed test point). These grids advance the understanding of how the sensor performs. Due to the circular shape of the mating hole on the sensor, it is expected that the results from each test will be a grid with a tight center circle of successful test points surrounded by unsuccessful test points. FIGURE 7 present results of random, horizontal, and vertical ordered tests with a 12x12 grid (for 144 total test points) and a stated key and sensor tolerance of 25 µm.

5. TESTING AND RESULTS

Preliminary sensor testing has been performed using this new automated test method. Testing was performed on two unique PVS's that differed in the material of a single internal component; one sensor's component is made from 3D printed plastic and the other sensor's component is made from machined stainless steel (SS). The PVS with the SS component was found to be of much higher precision than the PVS with the 3D printed plastic component.

Testing of the measurement test stand, along with capturing preliminary verification results of the PVS, began by conducting 55 total tests on a single PVS. These tests consisted of grid size XX and resolution YY. The order that the individual observations were collected varied across the 55 tests such that 19 tests were collected in random order, 18 tests in vertical order, and 18 tests in horizontal order. Ideally, it is expected that the successful sensor results would create a precise circular pattern, within the resolution of the test grid, with a diameter equal to the diameter of the key. All other areas of the test grid should result in

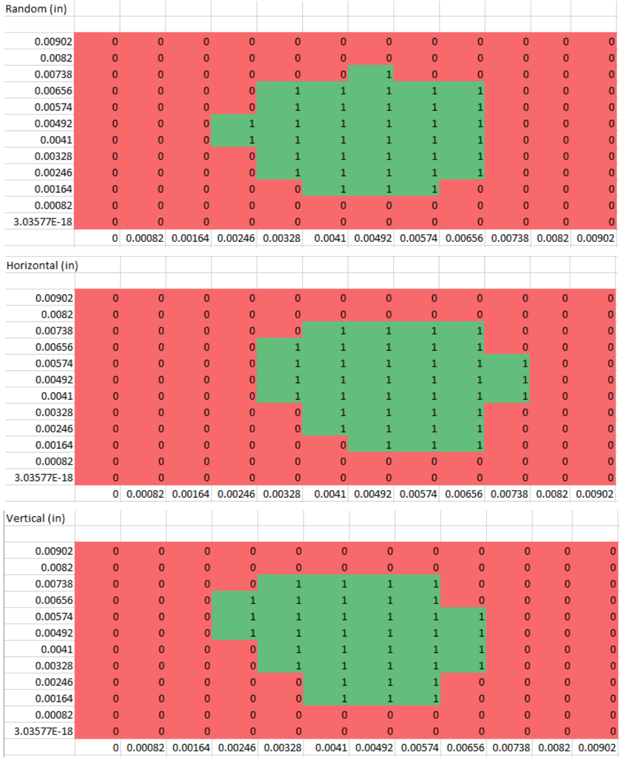
Weiss, Brian A.; Kaplan, Jared. "Assessment of a Novel Position Verification Sensor to Identify and Isolate Robot Workcell Health Degradation." Paper presented at ASME 2020 15th International Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020.



```
Currently testing point 238 of 397 at x=0.00850 inches and y=0.00550 inches.
The X motor position is 0.0085023622 inches.
The Y motor position is 0.0055015748 inches.
Process Unsuccessful

Currently testing point 328 of 396 at x=0.00500 inches and y=0.00800 inches.
The X motor position is 0.0050007874 inches.
The Y motor position is 0.0080023622 inches.
Process Unsuccessful

Currently testing point 215 of 395 at x=0.00700 inches and y=0.00500 inches.
The X motor position is 0.0070023622 inches.
The Y motor position is 0.0050011811 inches.
Process Successful

Currently testing point 140 of 394 at x=0.00950 inches and y=0.00300 inches.
The X motor position is 0.0095023622 inches.
The Y motor position is 0.0030000000 inches.
Process Unsuccessful
```

FIGURE 6. SAMPLE MATLAB OUTPUT

unsuccessful sensor results." The results of the testing done with this PVS deviated from the expected results. The circle of successful test observations encompassed most of the grid as opposed to the tight, expected center region. This led to the dissection of the first PVS revealing the internal 3D printed component. Given this material type is less precise than the machined SS, the SS PVS was then integrated into the measurement test stand. Fifteen total tests were run with the SS PVS: five random order, six vertical order, and four horizontal order; each with grid size XX and resolution YY. The grids created using this sensor were more closely aligned with what was originally anticipated; the grid was comprised of unsuccessful insertion tests, while the center of the grid was comprised of a tight, circular shape of successful tests. In total, 71 tests have been completed as of the preliminary development of this publication; 25 random order, 24 vertical order, and 22 horizontal order. These tests ranged in size from 12 observations (a very preliminary 3 x 4 grid to confirm the Matlab code) to 2500 observations. Test times ranged in duration from about five minutes to about 13 hours.



Random (in)												
0.00902	0	0	0	0	0	0	0	0	0	0	0	0
0.0082	0	0	0	0	0	0	0	0	0	0	0	0
0.00738	0	0	0	0	0	0	1	0	0	0	0	0
0.00656	0	0	0	0	1	1	1	1	1	0	0	0
0.00574	0	0	0	0	1	1	1	1	1	0	0	0
0.00492	0	0	0	1	1	1	1	1	1	0	0	0
0.0041	0	0	0	1	1	1	1	1	1	0	0	0
0.00328	0	0	0	0	1	1	1	1	1	0	0	0
0.00246	0	0	0	0	1	1	1	1	1	0	0	0
0.00164	0	0	0	0	0	1	1	1	0	0	0	0
0.00082	0	0	0	0	0	0	0	0	0	0	0	0
3.03577E-18	0	0	0	0	0	0	0	0	0	0	0	0
	0	0.00082	0.00164	0.00246	0.00328	0.0041	0.00492	0.00574	0.00656	0.00738	0.0082	0.00902

Horizontal (in)												
0.00902	0	0	0	0	0	0	0	0	0	0	0	0
0.0082	0	0	0	0	0	0	0	0	0	0	0	0
0.00738	0	0	0	0	0	1	1	1	1	0	0	0
0.00656	0	0	0	0	1	1	1	1	1	0	0	0
0.00574	0	0	0	0	1	1	1	1	1	1	0	0
0.00492	0	0	0	0	1	1	1	1	1	1	0	0
0.0041	0	0	0	0	1	1	1	1	1	1	0	0
0.00328	0	0	0	0	0	1	1	1	1	0	0	0
0.00246	0	0	0	0	0	1	1	1	1	0	0	0
0.00164	0	0	0	0	0	0	1	1	1	0	0	0
0.00082	0	0	0	0	0	0	0	0	0	0	0	0
3.03577E-18	0	0	0	0	0	0	0	0	0	0	0	0
	0	0.00082	0.00164	0.00246	0.00328	0.0041	0.00492	0.00574	0.00656	0.00738	0.0082	0.00902

Vertical (in)												
0.00902	0	0	0	0	0	0	0	0	0	0	0	0
0.0082	0	0	0	0	0	0	0	0	0	0	0	0
0.00738	0	0	0	0	1	1	1	1	0	0	0	0
0.00656	0	0	0	1	1	1	1	1	0	0	0	0
0.00574	0	0	0	1	1	1	1	1	1	0	0	0
0.00492	0	0	0	1	1	1	1	1	1	0	0	0
0.0041	0	0	0	0	1	1	1	1	1	0	0	0
0.00328	0	0	0	0	1	1	1	1	1	0	0	0
0.00246	0	0	0	0	0	1	1	1	0	0	0	0
0.00164	0	0	0	0	0	1	1	1	0	0	0	0
0.00082	0	0	0	0	0	0	0	0	0	0	0	0
3.03577E-18	0	0	0	0	0	0	0	0	0	0	0	0
	0	0.00082	0.00164	0.00246	0.00328	0.0041	0.00492	0.00574	0.00656	0.00738	0.0082	0.00902

FIGURE 7: TEST RESULTS FROM 12X12 GRID TEST OF A 0.001" TOLERANCE BETWEEN THE KEY AND SS PVS

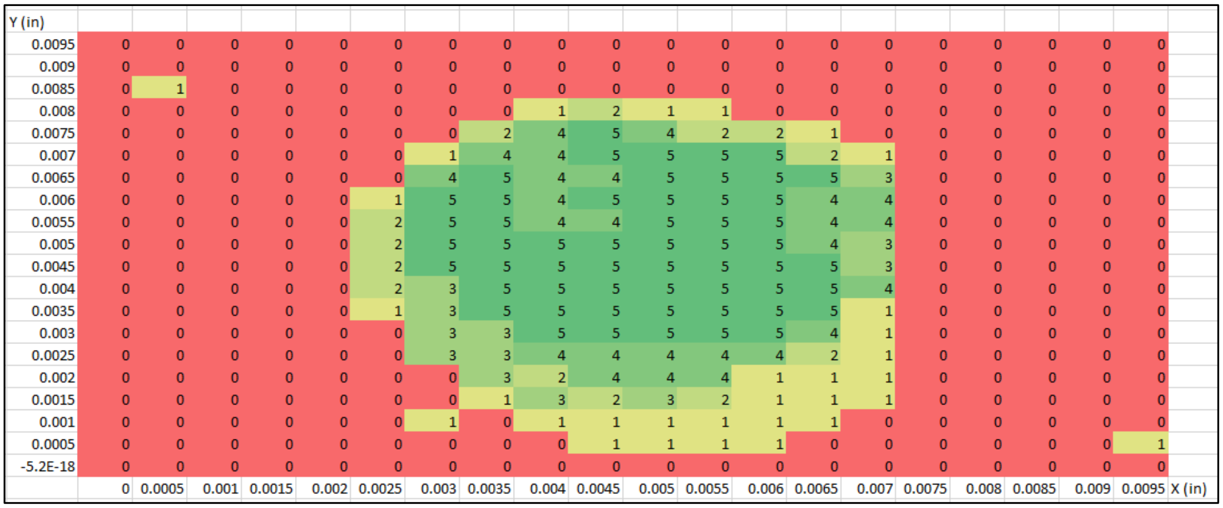
Some tests had unexpected anomalies; there were unsuccessful insertions in the middle of the grid where successful insertions were expected or successful insertions on the outside of the grid where failed insertions were expected. This was a very common problem that was faced throughout the entire testing process. FIGURE 7 presents three, 12 x 12 grid test results from random, horizontal, and vertical tests using the SS PVS that had a designed key/sensor tolerance of 25 μm. This means that the PVS opening is nominally 25 μm larger than the diameter of the key. For the specific tests that have been run and are discussed below, the PVS circular opening has a machined tolerance of 6.375 mm -0.0 to +12 μm while the tolerance of the cylinder key is 6.350 mm -0 to +5.08 μm. The non-binary numbers in FIGURE 7 represent the position, in millimeters, in the X and Y directions of travel from the X, Y coordinate origin of the key that is mated to the test stand.

Other tests included larger grid dimensions, such as 20 x 20 grids. FIGURE 8 presents a composite result of five separate, 20 x 20 tests that were performed. The results of the five tests were overlaid onto one another. The number provided at each individual point on the grid in FIGURE 8 is the number of total tests (out of 5) for which a successful insertion was observed. For example, a 3 represents a 60 % success rate where 3 insertions were successful while 2 insertions failed at the same point across five separate tests.

6. DISCUSSION

The material that is used to fabricate the two precision components of a single PVS was found to have a significant impact on the result of the test. 3D printed plastic yields less precise parts as compared to its machined SS counterpart. SS is a much stronger material that can be machined to tight tolerances making it a more advantageous material to use for the PVS's precision components. The PVS with the 3D printed plastic component has been removed from all further testing and workcell implementations. All PVS' in operation contain precision components made out of machined SS.

Further analysis uncovered additional deficiencies: 1) inconsistencies with the oscilloscope being used, 2) a defective automated linear stage, and 3) the geometry of the key. First, the oscilloscope that is used to measure the voltage from the PVS sensor would often emit a significant amount of electronic noise. This noise would cause the voltage values from the sensor to fluctuate to a degree that prevented the Matlab program from discerning between a successful and unsuccessful insertion test, leading to inaccurate results. Second, additional error stemmed from a breakdown of the linear, multi-axis stage that was used to move the pin to different points. The malfunction of the stage



Y (in)																					
0.0095	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.009	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.0085	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0.008	0	0	0	0	0	0	0	0	1	2	1	1	0	0	0	0	0	0	0	0	
0.0075	0	0	0	0	0	0	0	2	4	5	4	2	2	1	0	0	0	0	0	0	
0.007	0	0	0	0	0	0	1	4	4	5	5	5	5	2	1	0	0	0	0	0	
0.0065	0	0	0	0	0	0	4	5	4	4	5	5	5	5	3	0	0	0	0	0	
0.006	0	0	0	0	0	1	5	5	4	5	5	5	5	4	4	0	0	0	0	0	
0.0055	0	0	0	0	0	2	5	5	4	4	5	5	5	4	4	0	0	0	0	0	
0.005	0	0	0	0	0	2	5	5	5	5	5	5	5	4	3	0	0	0	0	0	
0.0045	0	0	0	0	0	2	5	5	5	5	5	5	5	5	3	0	0	0	0	0	
0.004	0	0	0	0	0	2	3	5	5	5	5	5	5	5	4	0	0	0	0	0	
0.0035	0	0	0	0	0	1	3	5	5	5	5	5	5	5	1	0	0	0	0	0	
0.003	0	0	0	0	0	0	3	3	5	5	5	5	5	4	1	0	0	0	0	0	
0.0025	0	0	0	0	0	0	3	3	4	4	4	4	4	2	1	0	0	0	0	0	
0.002	0	0	0	0	0	0	0	3	2	4	4	4	1	1	1	0	0	0	0	0	
0.0015	0	0	0	0	0	0	0	1	3	2	3	2	1	1	1	0	0	0	0	0	
0.001	0	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0	0	0	0	
0.0005	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	
-5.2E-18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	0	0.0005	0.001	0.0015	0.002	0.0025	0.003	0.0035	0.004	0.0045	0.005	0.0055	0.006	0.0065	0.007	0.0075	0.008	0.0085	0.009	0.0095	X (in)

FIGURE 8: COMPOSITE TEST RESULTS FROM 20X20 GRID TEST OF A 0.001" TOLERANCE BETWEEN THE KEY AND PVS

limited its range of motion and prevented it from moving to the desired location when commanded to do so by the Matlab program, which significantly impacted the results of some of the tests. The malfunctioning stage was replaced with a new, identical stage and testing continued. Lastly, it is important to note that the keys inserted into the PVS are COTS high precision cylinders with notionally known dimensions. Unfortunately, the cylinder contains a very small chamfer at the cylindrical edges. This chamfer feature likely resulted in several additional boundary test points being successful, as opposed to failing, since this chamfer would allow the pin to 'slide in' if the pin was slightly out of tolerance.

The impacts of these deficiencies were obvious for several tests whose results have been discarded entirely, however, the deficiencies may have also affected tests that have been considered "good". The results of the insertion process at one or more points in these "good" tests could have been changed due to these shortcomings. The malfunctioning stage issue has been resolved and will have minimal effect on future testing. The oscilloscope issue is still being investigated in order to find a solution to the problem, which may involve changing oscilloscope settings or changing the method used to measure sensor voltage. Regarding the COTS keys with slight chamfer, these keys are still in use until a viable replacement is determined. One potential solution is to machine custom keys of similarly tight tolerances with no chamfer, yet there will have to be some type of finishing operation performed on the edges of the cylinders to remove any sharp edges. This would provide a natural, yet slight chamfer. It's possible that the long-term solution is to continue to use the COTS keys and appropriately adjust the subsequent results. For example, a known, designed tolerance of 25 µm between the COTS key and the PVS could translate into successful insertion test results demonstrating a 63.5 µm tolerance.

The optimal test (i.e., size of grid, spacing between points, and optimal point selection process) for verification of the PVS is still being determined. Ideally, this verification test would provide enough information on the behavior and integrity of the PVS to generate a quantifiable confidence in its measurement capability; and be performed within a reasonable amount of time. The grid sizes that have been used so far have ranged from 3 x 4 to 50 x 50 and the resolution of the spacing between points has ranged from 100 µm to 5 µm. The results of the tests performed at each of these sizes and resolutions varied between the three different selection methods. The three order of observations methods (random, vertical, and horizontal) were applied at each of grid sizes and resolutions considered. More detailed grids generate more data on the PVS and its behavior, until a capacity is achieved where no new information can be gathered.

Overall, the preliminary test efforts proved insightful, both in terms of developing an initial understanding of the PVS' performance and the capability of this specific verification test method. From preliminary test results, including several composites that have been produced, it's evident that a key and PVS pairing with a designed 25 µm clearance will have successful insertion tests greater than 25 µm due to the factors discussed earlier in this section.

7. FUTURE EFFORTS

More testing is still required before reaching the end goal of releasing the sensor to industry for ubiquitous use. All the testing that has been performed so far has been done using PVS' with the specific clearance of 25 µm relative to the key. Future efforts will involve more testing of sensors with a key clearance of 25 µm, as well as other sensors with larger key clearances (e.g. 50 µm, 100 µm, 254 µm etc.). With additional testing and improvements, the PVS can be verified for implementation and use in manufacturing environments. The ultimate goal is for the *Identification and Isolation of Robot Workcell Accuracy Degradation* test method and the PVS to be used by industry to monitor and respond to changes in the health of their robot workcells.

The Matlab program used in the testing process will also be updated to add three new functionalities: a graphical user interface (GUI), an option to test grid boundary conditions (the regions where the test results change from successful to unsuccessful), and an expanded analysis and visualization capability within the Matlab platform. The GUI will make the testing process more user friendly for personnel who are not familiar with the program. The GUI is currently in the very early stages of development. The boundary testing functionality will allow the user to move the stage to the boundary points of the center circle and increment outwards by very small amounts (less than the distance between points) until reaching a point where the key and sensor are no longer able to mate. This will give the user a more accurate distribution, along with a more accurate representation of how the PVS behaves and crystallize the PVS' boundaries in terms of where successful tests become unsuccessful. This testing approach will gather new information about the PVS, such as a more accurate representation of the room for error that is acceptable to successfully insert the key into the PVS.

The current PVS that is being used in the test method is a binary sensor. It can communicate to users that an element along the workcell's kinematic chain either passes or fails the test, indicating whether the workcell is healthy or not. A new PVS is being developed with the ability to communicate a greater granularity of workcell health. For example, the workcell is healthy and no maintenance is needed, the workcell health is degrading but it is not affecting part quality, or the workcell is unhealthy and part quality is negatively impacted (this scenario would require immediate attention). This new PVS design expands upon the initial iteration adding another inner button and nesting a second outer circle of specific diameter. Additional details on the sensor are discussed in [28]. This intelligence would better enable manufacturers to optimize their maintenance efforts and schedules. This new generation of PVS will also require testing and verification. Future efforts will involve using the same, or a similar, process to test and verify this new sensor for release to manufacturers.

Lastly, it is intended that the data from the execution of the *Identification and Isolation of Robot Workcell Accuracy Degradation* will augment process and equipment intelligence with respect to health and maintenance activities. The PVS' binary output (pass or fail) of key elements along a robot workcell's kinematic chain, could be coupled with one or more of the data types presented in the BACKGROUND section to enhance overall maintenance intelligence of a manufacturing operation or speed deeper troubleshooting of a workcell. Once the PVS is further verified through additional testing, it's continued use within manufacturing facilities will present opportunities to capture binary test method data with real manufacturing data. This data will be correlated to better understand degradation trends and relationships among data types. Identifying redundant or inconsequential data can have a substantial impact on future data collection and analysis efforts. Ideally, manufacturers will only capture data from specific sources to acquire targeted intelligence leading to decisive and cost-effective maintenance actions.

NIST DISCLAIMER

The views and opinions expressed herein do not necessarily state or reflect those of NIST. Certain commercial entities, equipment, or materials may be identified in this document to illustrate a point or concept. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] Michalos, G., Makris, S., Papakostas, N., Mourtzis, D., and Chryssolouris, G., 2010, "Automotive assembly technologies review: challenges and outlook for a flexible and adaptive approach," CIRP Journal of Manufacturing Science and Technology, 2(2), pp. 81-91.

[2] Muller, R., Esser, M., and Vette, M., 2013, "Reconfigurable handling systems as an enabler for large components in mass customized production," Journal of Intelligent Manufacturing, 24(5), pp. 977-990.

[3] Jovane, F., Koren, Y., and Boer, C. R., 2003, "Present and future of flexible automation: Towards new paradigms," Cirp Annals-Manufacturing Technology, 52(2), pp. 543-560.

[4] Marvel, J. A., and Newman, W. S., "Accelerating robotic assembly parameter optimization through the generation of internal models," Proc. Technologies for Practical Robot Applications, 2009. TePRA 2009. IEEE International Conference on, IEEE, pp. 42-47.

[5] Qiao, G., and Weiss, B. A., 2016, "Advancing Measurement Science to Assess Monitoring, Diagnostics, and Prognostics for Manufacturing Robotics," Int J Progn Health Manag, 7(Spec Iss on Smart Manufacturing PHM), p. 13.

[6] Park, C., and Park, K., "Design and kinematics analysis of dual arm robot manipulator for precision assembly," Proc. Industrial Informatics, 2008. INDIN 2008. 6th IEEE International Conference on, IEEE, pp. 430-435.

[7] Baglee, D., and Jantunen, E., "Can equipment failure modes support the use of a condition based maintenance strategy?," Proc. 3rd International Conference on Through-Life Engineering Services, TESConf 2014, Elsevier, pp. 87-91.

[8] Ly, C., Tom, K., Byington, C. S., Patrick, R., and Vachtsevanos, G. J., "Fault diagnosis and failure prognosis for engineering systems: A global perspective," Proc. 2009 IEEE International Conference on Automation Science and Engineering, CASE 2009, IEEE Computer Society, pp. 108-115.

[9] Williams, R., Banner, J., Knowles, I., Dube, M., Natishan, M., and Pecht, M., 1998, "An investigation of 'cannot duplicate' failures," Quality and Reliability Engineering International, 14(5), pp. 331-337.

[10] Jin, X., Siegel, D., Weiss, B. A., Gamel, E., Wang, W., Lee, J., and Ni, J., 2016, "The present status and future growth of maintenance in US manufacturing: results from a pilot survey," Manuf Rev (Les Ulis), 3, p. 10.

[11] Helu, M., and Weiss, B. A., "The current state of sensing, health management, and control for small-to-medium-szed manufacturers," Proc. ASME 2016 Manufacturing Science and Engineering Conference, MSEC2016.

[12] Weiss, B. A., Vogl, G. W., Helu, M., Qiao, G., Pellegrino, J., Justiniano, M., and Raghunathan, A., 2015, "Measurement Science for Prognostics and Health Management for Smart Manufacturing Systems: Key Findings from a Roadmapping Workshop," Annual Conference of the Prognostics and Health Management Society 2015, P. Society, ed., PHM Society, Coronado, CA, p. 11.

[13] Pellegrino, J., Justiniano, M., Raghunathan, A., and Weiss, B. A., 2016, "Measurement Science Roadmap for Prognostics and Health Management for Smart Manufacturing Systems," NIST Advanced Manufacturing Seriess (AMS).

[14] Helu, M., and Hedberg, T., 2015, "Enabling Smart Manufacturing Research and Development using a Product Lifecycle Test Bed," 43rd North American Manufacturing Research Conference, Namrc 43, 1, pp. 86-97.

[15] Klinger, A. S., and Weiss, B. A., 2018, "Robotic Work Cell Test Bed to Support Measurement Science for PHM," 2018 ASME Manufacturing Science and Engineering Conference (MSEC), American Society of Mechanical Engineers (ASME), College Station, Texas.

[16] Klinger, A., and Weiss, B. A., 2018, "Examining Workcell Kinematic Chains to Identify Sources of Positioning Degradation," Annual Conference of the PHM SocietyPhiladelphia, Pennsylvania, p. 9.

[17] Jin, X., Weiss, B. A., Siegel, D., and Lee, J., 2016, "Present Status and Future Growth of Advanced Maintenance Technology and Strategy in US Manufacturing," Int J Progn Health Manag, 7(Spec Iss on Smart Manufacturing PHM), p. 012.

[18] Brundage, M. P., Sexton, T., Hodkiewicz, M., Morris, K. C., Arinez, J., Ameri, F., Ni, J., and Xiao, G., 2019, "Where do we start? Guidance for technology implementation in maintenance management for manufacturing," Journal of Manufacturing Science and Engineering, 141(9).

[19] Baybutt, M., Minnella, C., Ginart, A., Kalgren, P. W., and Roemer, M. J., "Improving digital system diagnostics through Prognostic and Health Management (PHM) technology," Proc. 42nd Annual IEEE AUTOTESTCON Conference 2007, Institute of Electrical and Electronics Engineers Inc., pp. 537-546.

[20] Holland, S. W., Barajas, L. G., Salman, M., and Zhang, Y., "PHM for Automotive Manufacturing & Vehicle Applications," Proc. Prognostics & Health Management Conference.

[21] Tsui, K. L., Chen, N., Zhou, Q., Hai, Y. Z., and Wang, W. B., 2015, "Prognostics and Health Management: A Review on Data Driven Approaches," Mathematical Problems in Engineering, 2015, pp. 1-17.

[22] Vogl, G. W., Weiss, B. A., and Donmez, M. A., 2014, "Standards Related to Prognostics and Health Management (PHM) for Manufacturing," No. NISTIR 8012, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA.

[23] Vogl, G. W., Weiss, B. A., and Helu, M., 2019, "A review of diagnostic and prognostic capabilities and best practices for manufacturing," Journal of Intelligent Manufacturing, 30(1), pp. 79-95.

[24] Sharp, M., Brundage, M. P., Sprock, T., and Weiss, B. A., "Selecting Optimal Data for Creating Informed Maintenance Decisions in a Manufacturing Environment," Proc. Model-Based Enterprise Summit 2019.

[25] Pan, M. C., Van Brussel, H., and Sas, P., 1998, "Intelligent joint fault diagnosis of industrial robots," Mechanical Systems and Signal Processing, 12(4), pp. 571-588.

[26] Qiao, G., Schlenoff, C. I., and Weiss, B. A., 2017, "Quick Positional Health Assessment for Industrial Robot Prognostics and Health Management (PHM)," IEEE International Conference on Robotics and Automation 2017Singapore, p. 6.

[27] Weiss, B. A., and Klinger, A. S., "Identification of Industrial Robot Arm Work Cell Use Cases and a Test Bed to Promote Monitoring, Diagnostic, and Prognostic Technologies," Proc. 2017 Annual Conference of the Prognostics and Health Management (PHM) Society, PHM Society, p. 9.

[28] Weiss, B. A., 2019, "Developing Measurement Science to Verify and Validate the Identification of Robot Workcell Degradation," ASME 2019 14th International Manufacturing Science and Engineering Conference (MSEC2019), ASME, Erie, Pennsylvania, p. 10.

UNDERSTANDING AND EVALUATING NAIVE DIAGNOSTICS ALGORITHMS APPLICABLE IN MULTISTAGE MANUFACTURING FROM A RISK MANAGEMENT PERSPECTIVE

Mehdi Dadfarnia, Michael Sharp, Timothy Sprock
National Institute of Standards and Technology
Gaithersburg, MD 20899

ABSTRACT

The world has entered a state of unprecedented access to machine intelligence algorithms, where the ease of deployment has created a scenario where nearly every facet of life and industry has been affected by AI. Especially within industry, where the options for enacting AI systems are wide and varied, the choice of which system will work best for a given application can be daunting. Understanding when, where, and why to apply a particular algorithm can provide competitive advantage on effectiveness as well as greater trust and justification when using the algorithms' outputs. This paper examines multistage manufacturing processes, where system complexity can greatly influence the burden of creating custom tailored monitoring solutions. Such barriers have encouraged many manufacturing small and medium enterprises (SME) to look towards generic 'black box' commercial software solutions, although they may lack the sufficient expertise to objectively determine which product best meets their requirements. Some of the considerations faced by SMEs are identifying tools that can successfully be deployed alongside a potential lack of sensor coverage and/or the desire for rapid system reconfiguration to accommodate smaller custom batch production sizes. In these environments, detailed analytics-based solutions are often not feasible for production equipment monitoring. This paper provides a procedure for assessing the suitability of various tools or algorithms used to evaluate production process performance based on product quality output. This paper also presents a preliminary comparative example study of several algorithms to demonstrate this process and evaluate the selected algorithms.

Keywords: Manufacturing simulation, problem diagnosis, fault isolation, evaluating algorithms.

1. INTRODUCTION

Manufacturing is a highly competitive industry where every decision should be qualified to ensure both effectiveness as well as a solid return on investment. A common choice faced by manufactures is the decision of if, when, and how to monitor both the quality of their product and the effectiveness of the machines used in the processes. Due to limits in resources, especially for SME manufacturers, the availability of information sources such as sensors to monitor individual machine effectiveness may be severely restricted or otherwise unsuitable for analysis. However, the two forms of information available to nearly any sized manufacturer are the end part quality and process path used to create each individual part.

Tracking part quality information and quickly identifying sources of problems has become particularly important in agile multistage manufacturing facilities where small batch sizes and rapid reconfiguration are vital to maintain a competitive edge. Rapidly changing system dynamics can exacerbate and propagate problems in machine performance across multiple product sets, costing thousands of dollars if not quickly identified and managed. Selecting proper tools and methods for monitoring system performance is a significant decision that can have long-term implications for management and factory operators who will have to interpret and interact with any deployed monitoring system. Being able to understand and justify decisions regarding the selection of a monitoring system can increase confidence in that system and help users understand any risk associated with its use on a factory floor.

Past research has focused on utilizing part quality data and part process path information to aid in the determination of problematic equipment or process links as the basis for maintenance activities [1]. Some of these activities are very specific, such as using historical part quality degradation to augment and improve linear system dynamics models to predict metrics such as tool wear [2]. The effort and effectiveness of each technique can vary widely between applications. This work seeks to provide a comparative analysis of several popular fault or problem isolation algorithms and to explore the general areas where they are most and least effective at identifying causes of part quality degradation. From this, a workflow is developed for testing the range of applicability of any algorithm

Dadfarnia, Mehdi; Sharp, Michael; Sprock, Timothy. "Understanding and Evaluating Naive Diagnostics Algorithms Applicable in Multistage Manufacturing from a Risk Management Perspective." Paper presented at ASME Manufacturing Science and Engineering Conference 2020 (MSEC 2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

as well as a basic guide for determining suitability of tested algorithms for various system setups. This paper is not meant to be an exhaustive comparison of all popular methods of multistage manufacturing monitoring, instead it is directed at developing an environment which can be used to critically evaluate tools and methods in a manner which allows for objective comparative assessment.

1.1. Background and Motivation

Stream of variance modeling has been a staple of multistage manufacturing since it was introduced in the 1990s [3]. Developed because rigid body assumptions do not always hold through production, stream of variance modeling recognizes that multistage processes can add compounding errors through both parts handling and machining. Some work focused directly on diagnosing fixture variation in multistage manufacturing processes [4] and led to further investigations of machining errors through explicit system modeling, such as via linear state space models [5]. Later work in statistical process control with linear state space models was able to utilize probability and hypothesis testing to capture faulty elements within a process [6]. Intuitively, many of these techniques are sensitive to the measurements and recordings used as inputs to the monitoring algorithm [7], meaning that they require significant sensing capabilities throughout the system. Other barriers to correctly implementing stream of variance algorithms include creating an accurate representation of features, selecting an optimal parts sampling criteria for inspection, and developing an adequate level of detail in modeling of the possible process faults [8].

Although these and related methodologies can provide important and accurate information regarding the propagation of errors as well as their initial incident location, explicit system models tend to require an advanced level of expertise, both with the system dynamics and the algorithms themselves, to properly implement. Additionally, many monitoring strategies relying on explicit system modeling become impractical for increasingly complex systems. To combat this, Huang et al. (2002) suggest simplifications and substitutions to mitigate this problem of growing complexity [9]. Many of these simplifications are based around the notion that there are certain configurations which will obfuscate areas from pinpoint diagnosis by simple virtue of their design. Zhou et al. (2003) understood that areas of obfuscation within a system, while unavoidable in some situations, would be strictly undesirable from a monitoring standpoint and proposed ways to quantify them [10]. In general terms, areas within a system become indistinguishable from one another if no unique information is generated between any subset of elements (within that area of the system).

Even with methods for simplifying the system model, often smaller enterprises do not develop monitoring tools that require explicit system models because the system dynamics change too rapidly to make any explicit modeling a practicality. This has driven investigations into less analytically-explicit monitoring algorithms. Further, due to the comparatively low amount of "stable data" produced by reconfigurable systems, practitioners seek algorithms that can operate with minimum input. One approach relies on using historical data to develop patterns of expected behavior from the machines based on post-production (or intermittent production) product quality reports, then comparing current behavior to develop diagnostic information. For example, methods for comparing expected distribution curves [11] and data- data-driven techniques for variation reduction [12] are gaining acceptance in multistage manufacturing.

Ultimately increasing data integration in automation and manufacturing depends on strong algorithms supporting human decision-making [13]. The strength of these algorithms depends on their performance evaluation and applicability. Evaluating the wide variety of algorithms in development and in practice has been inconsistent and traditionally reliant on the expertise levels of the developers or practitioners involved in the algorithm deployment. Most evaluations will stop if there exists a comparative analysis against one other algorithm on a limited set of system scenarios or data. This is due in part to the large time investment required to set up a large comparative study, but also due to a lack of understanding about how to set up such a study. In order to best evaluate an algorithm's suitability for a system, a majority of potentially disruptive scenarios must be considered. Regardless of the method or selection and availability of input data, the procedure for qualifying an algorithm or tool on a system must consider a majority of relevant edge cases as well as the nominal expected scenarios.

This work explores evaluating process fault or problem diagnostic tools with a very limited set of input data, specifically process path and part quality. That does not preclude extension of these procedures to more extensive algorithm testing. The choice to focus on algorithms that utilize limited information was motivated to both highlight the procedure on a simple comparative case with standard input parameters and to help shed light on real decision-making problems faced by SMEs. This case study will examine some of the 'black-box' solutions that are being applied to this problem. These include probabilistic statistical algorithms, gradient descent, genetic algorithms, and neural networks. These solutions will be framed to evaluate unsophisticated applications similar to that as may be identified by an SME seeking low-cost solutions to integrating new monitoring programs.

2. METHODOLOGY

This paper utilizes a general workflow for testing various fault or problem isolation algorithms. Defined or obtained series of system configurations will serve as the validation environment for obtaining exemplar data to investigate representations of a wide range of normal and off-normal operations. Special emphasis should be placed on investigating scenarios for both configurations and conditions that are reasonably expected to exhibit themselves in practice and would be antagonistic towards the algorithms being evaluated. In most multistage manufacturing systems, it is unreasonable to attempt to investigate all possible scenarios; part of this work is meant to help guide and highlight the process of determining high-risk edge cases that will provide the most pertinent information regarding the algorithm being evaluated. The final steps of the evaluation procedure are to apply the selected algorithms and measure them via metrics most suited to the end goals of the production line and, if necessary, iterate through the antagonistic scenarios to develop confidence in the coverage and performance of the algorithms.

2.1. Selecting Testing Scenarios

To show the viability and effectiveness of this workflow, a simulator was developed that can represent a variety of multistage manufacturing systems and conditions. For an actual manufacturing facility, it may be sufficient to simulate the actual or possible configurations for that specific production floor, possibly augmented with available historical data. This paper focuses on cases where the particular system that an algorithm will be tested on is not known beforehand. This additionally addresses the broader research perspective on which types of systems an algorithm is useful for. Testing various configurations follows from understanding that system configuration can profoundly impact the performance of the manufacturing process [14], and thus any monitoring tools deployed on them.

Each testing scenario consists of two major aspects: the system configuration and the condition. The configuration relates to how the equipment is connected and the process flow. The condition conveys the health of machines, production rates, etc. In many systems, the large number of possible system configurations and conditions makes it impractical to test every single one. Therefore, the testing methodology focuses on three primary classes of scenarios: nominal, high-risk, and those that are antagonistic towards the algorithm being tested. Some scenarios may fall into multiple classes, but a proper set of testing scenarios should provide sufficient coverage representing these classes as evenly as is practical. Finally, each scenario has a likelihood that gives a relative expectation of how often that scenario would be expected to occur in the lifetime of the system.

A major driver in selecting scenarios is the potential consequences of that scenario, both good and bad. Understanding the potential outcomes of combinations of missed fault alarms, false fault alarms, as well as correct alarms all can help drive the selection of scenarios. Coupled with the associated frequency, a potential risk for each scenario can be developed and used to identify important test cases. The risk metric (frequency * consequences) should be evaluated in terms most suited to the application. In a manufacturing setting, for example, risk can be measured as production loss per hour. Once acceptable levels of risk are established, any scenario which falls below that can be safely ignored.

2.1.1. Low-Risk High-Frequency and High-Risk Scenarios

The scenarios that should be developed first are those capturing the most common configurations and expected conditions of the factory environment. Even if the risk of these scenarios is fairly low, the fact that they are the most common scenarios necessitates their evaluation for any potential monitoring algorithm. These include scenarios such as a single faulty machine that does not stop operations. One occasionally overlooked scenario that should always be included is the most common configuration(s) with fault-free, nominal conditions. This is critical to help characterize the false positive rates of any diagnostic algorithms and establish a risk based on false alarms.

The second priority scenarios are those that may or may not be common but present a high-risk factor if the algorithm is unable to correctly identify the source of incipient defects or problems. High risk can come from both false positives and false negatives in terms of identification and must be evaluated for different combinations of both.

2.1.2. Antagonistic Scenarios

Antagonistic scenarios are less intuitive because they are typically more dependent on the algorithm than the physical process, but they are important to understand when interpreting evaluation testing. Antagonistic scenarios are configurations and conditions for which an algorithm is expected to perform poorly based on the assumptions and capabilities of that particular algorithm. These types of scenarios may not be known prior to the evaluation of the algorithm, but if a set of poor performing scenarios begins to develop, any identifiable commonalities can be used to group the full set of test scenarios and add extra examples to those groups if needed.

Once a set of antagonistic scenarios is developed, the existing suite of test scenarios can be checked for how commonly those antagonistic qualities occur. If they are prevalent, especially in high risk scenarios, there may not be a need to continue evaluating that algorithm, because it would be expected to perform poorly in these high-risk scenarios. However, if there are not many scenarios with these common traits, it is prudent to construct more such scenarios and evaluate if these have a high enough cumulative risk to affect the decision of accepting the algorithm.

2.2. Obtain Evaluation Data

Simulations and/or historic data will be required to perform these tests. Where available, simulators would generally be desirable to augment the available scenarios that can be evaluated. Real systems with huge backlogs of data can be used as an initial set of testing scenarios, but it is unlikely that real systems have substantial amounts of high-risk data under the array of faulted conditions that would be desired to fully evaluate a potential scenario. The flexibility of a simulator, augmenting any available real scenario data, allows for better exploration of high risk and antagonistic scenarios. Existing data or logs of activities can also be used to help develop the associated frequencies and consequences of any scenario. Maintenance and production logs may be a good source of this information.

2.3. Evaluate Algorithms

Once the test scenarios and data has been obtained, the algorithms should be streamed onto the data as they would receive it in an actual production environment. For example, if there is not a live stream of the product quality available in the plant, then the algorithm should be provided batch style updates with corresponding time stamps and other relevant meta-data. The processing time of the algorithm under evaluation should also be noted at this time. If the routine takes longer to process than the update rate for the system, special considerations and accommodations must be made, such as artificially slowing the input rate. If this cannot be done, or the accommodations are too severe to work at scale, the algorithm may be deemed unusable without further testing.

2.4. Representing Key Algorithm Performance Indicators

When evaluating algorithms, it is important to select performance indicators that not only reflect the performance of the algorithms but do it in a way that is relatable to the end goals of the monitoring algorithms. For most diagnostic isolation problems, the most important metrics are those that directly relate to negative consequences: false positives - identifying a fault where there isn't one, false negatives - failing to identify a fault, and circumstances that produce both. Because the repercussions of these three outcomes are generally significantly different and may even relate to specific equipment within the system, it is most appropriate and convenient to record the performance of the algorithms as a series of confusion matrices, one for each testing scenario. The important aspect for a risk-based evaluation is that each testing scenario have a numeric representation of the rate of negative consequences that can be translated into a probability during the final evaluation.

2.5. Summary

Below is a summary of the algorithm evaluation process presented in this section. The methods and selection criteria presented in this section are not intended to be interpreted as the only, or best possible criteria for every case. Instead they are described as a possible set that would be applicable in most cases and are the methods used in this work. The next section will describe the specifics of this work as applied to the developed test cases. The algorithms chosen for evaluation are also described, followed by the general outcomes of that investigation.

1. Define Testing Environment
2. Select Testing Scenarios
   a. Nominal and High-Risk Scenarios
   b. Antagonistic Scenarios
3. Obtain Data Covering Selected Scenario
4. Evaluate Selected Algorithm(s)
   a. Check Usability / Operational Concerns
   b. Evaluate Performance
5. Repeat as needed to discover edge cases with poor performance

3. PROOF OF CONCEPT CASE STUDY

This case study highlights the methodology for evaluating a specific class of monitoring algorithms that may be used to determine potential locations of induced damage or defects in products output from multistage manufacturing processes. The scenarios and configurations presented are selected to allow broad level performance investigation of algorithms on a potentially unknown system. A simulator was created as part of this work to rapidly create arbitrary scenarios generating the required end-of-line part quality and part process production path data. This data is then used to evaluate multistage manufacturing system diagnostic algorithms. The setup of the simulator includes the manufacturing system configuration (section 3.1), test case scenarios that characterize nominal and edge cases faced by the system (section 3.2), and diagnostic algorithms applied to analyze the scenarios (section 3.3). This evaluation is presented as preliminary work and additional work is needed to completely characterize the algorithms under evaluation.

3.1. Define Testing Environment

Figure 1 shows an example of different multistage manufacturing paths to produce a part, given a limited number of paths and available machines. Machines can be used across multiple paths (process plans), subject to timing restrictions, resulting in a directed graph representation of product production. Here we assume that it is possible for different production paths to use the same machine, but not possible for any path to use a particular machine more than once (reentrant flows). The subset of machines within each process path is assigned, either randomly or manually, at the beginning of each scenario and held static for the duration of the test.

The type of manufactured product is unspecified during the simulation, but it is assumed that quality metrics of any produced parts are scalable to some equivalent metric. Each part is given a single score (*Q_part***)** to indicate its build quality at the end of its sequence (end-of-line part quality). End quality assessments (*Q_part*) are scaled to a percentage of acceptability, where 100 % is perfect and anything below 0 % is considered lost product. This score relies on the added value from each machine that the part interacts with in its production path. Added quality value to the part is uniform across all machines unless the part interacts with a machine that underwent a sudden degradation or failure. In this case, the degraded machine will subtract from the part's accumulated value. Currently, part quality inspection is limited to the end of each production path.

Although simplified and abstract, this system configuration simulator exhibits key behaviors of real-world production settings, including resource constraints, a limited number of production path setups, and the limited number of available machines for production. Further development of the system configuration in future studies will include constraints on machine-ordering in a path (i.e., having rules such as "machine #6 may never be used before machine #3") and available

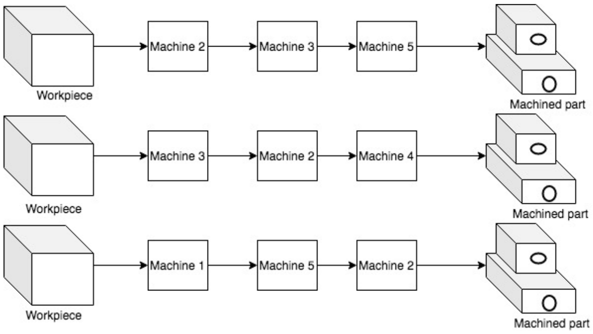


Figure 1. An example of three, unique three-stage machining paths to produce a part given five available machines.

machines categorized by different types, with constraints on the number of each machine type required to build a part. The machines may also have a more expansive feature space that determines each machine's added value to a part's quality.

3.2. Select Testing Scenarios

The above setup provides a framework to simulate different scenarios that could be used to evaluate multistage manufacturing system diagnostic algorithms that attempt to identify causes for part quality degradation. The focus in this paper is on part quality degradation due to sudden machine degradation or failure. The simulator receives specifications that after a certain number of products, a subset of the total set of available machines will degrade their added manufacturing values to the quality of parts that interact with it. As the quality for each part relies on the accumulation of the added values from all machines that are in the part's production path, the paths that have one or more of these degraded machines will output parts with lower end-of-line part quality values. Figure 2 illustrates a temporal view of the part quality outputs for an example test case scenario. Note that although the scalar part quality values are arbitrary, there is a drop-in value for parts that have been processed by some of the paths (that include the machine degradation at *t*=200 s).

The inputs into the configuration of the simulator are the number of machines on the "production floor", the number of processing steps each item needs to be created, the number of production paths or lines that a part can be created on, and the rate at which those parts are made. These inputs define a system of machines and paths involved in the multistage manufacturing process to produce each part. These key manufacturing parameters that go into the system configuration are summarized in Table 1. Presented in this paper are two preliminary scenarios defined by the values given in Table 1.

Test case scenarios are built to evaluate how well diagnostic algorithms identify degraded machines. In particular, it is interesting to evaluate the response of these diagnostic algorithms to different numbers of machines degraded at a time. Using the number of degraded machines as the main experimental factor in test case scenarios provides insight into the usability and accuracy of each tested algorithm for different numbers of degraded machines. These test cases are implemented on the two system configurations in Table 1.

Table 1. System Configuration Parameters

Description	System #1	System #2
Total number of available machines	10	12
Machines in a production path	4	6
Number of unique production paths	13	20
Part production rate (part per second)	1	1

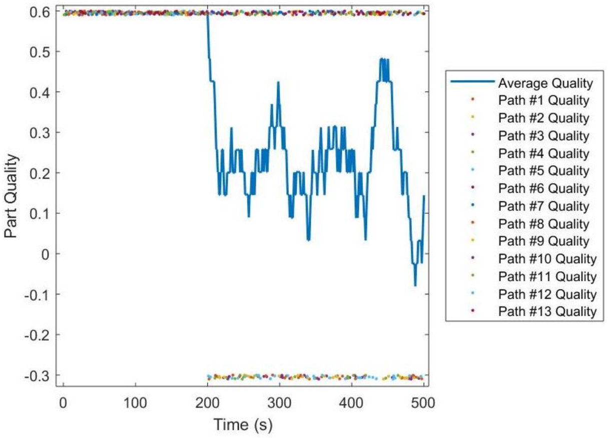


Figure 2. Example of a temporal view of part quality production. Machine degradation begins at t=200 s. Subsequently, paths that include the degraded machine produce lower-quality products.

For the first system configuration example, where there are 13 unique production paths and 4 machines (selected from an available 10) in each path, there are three levels to the experiment factor: one, two, or three machines are degraded. A maximum of three degraded machines is selected because it reaches sufficient coverage over scenarios where most of the machines in some production paths (3 of 4) are producing undesirable results.

Ten experiments, or test case scenarios, are performed for each of these three machine degradation levels. Each of these experiments is performed with different randomly-selected permutations of degraded machines. Ten is the maximum number of permutations when only one machine degradation occurs out of ten available machines. The number of possible permutations for two or three machine degradations is much higher. For the sake of brevity and producing initial results, the other machine degradation levels were limited to ten test case scenarios as well.

Table 2. Test Case Scenario Setup

	System #1	System #2
Number of Factor Levels	3	4
Number of Experiments Per Level	10	12
Total Number of Test Case Scenarios	30	48
Degradation Start Time (s)	200	200
Number of parts (stopping time)	500	800

Test case scenarios are similarly designed on the second system configuration example, where there are 20 unique production paths and 6 machines (out of an available 12) in each path. Since the scope of this example's multistage manufacturing system is larger, there are instead four levels to the experiment factor (a maximum of four degraded machines) and 12 experiments with different permutations at each level (12 is the maximum number of permutations when only one of 12 available machines degrade). Table 2 summarizes the test case scenario setup for each of the system configurations in Table 1.

The simulator generates the scenarios and stores the machines that degrade in each test case. The sequence of machines in the paths of each system configuration is held constant against the different machine degradation schemes from the test cases. This allows for applying diagnostic algorithms on each of these test case scenarios to compare their prediction of degraded machines against the actual machine failures as well as to evaluate their applicability to different machine degradation scenarios (i.e. different numbers of machines that degrade at a time).

The system configuration and test case scenario setups also enable the derivation of different properties from the system configurations. These properties may include the frequency of degraded machines within all the production paths or the amount of information that can be distinguished from the machines traversed in each path. This gives the opportunity to manipulate the properties and observe the prediction response of diagnostic algorithms to changes in different properties in the system configuration. This provides insight into the applicability and limitations of the diagnostic algorithms when there are changes in the system configuration. Discovering these insights requires changing the test case scenario setup to use the derived properties as experimental factors in a more extensive, robustly-designed experimental methods [15]. Deriving these properties and designing experiments around them are topics intended to be explored in future studies. Also, this paper has only been looking into the case of sudden machine failures. Future studies will expand to look at different types of machine failure modes, such as observing when diagnostic algorithms identify more gradual machine degradations.

3.3. Diagnostic Algorithms Application Setup

Five different diagnostic algorithms have been selected for evaluation and applied on each of the test scenarios for this work. Each algorithm makes use of only two sources of information: the end-of-production-line part quality, and the part process production path. Testing different algorithms allows to identify strengths and weaknesses of each one with respect to the different system configurations, amounts of machine induced degradation, and locations of induced damage. The selected algorithms are only a small random sampling of potentially useful methods for isolating induced faults on a multistage production line. The following describe the five algorithms tested.

3.3.1. Probabilistic Statistical Algorithms

One of the more intuitive mechanisms for performing diagnostic isolation is to evaluate the probability or likelihood that a given machine is producing a defective part based directly on the observed outputs corresponding to that process element. This could be framed under the guise of temporal difference reinforcement learning, a relative to Q learning. Presented below are two algorithms that utilize the intuitive nature of the process to create logical evaluations of the relative probability any given machine (process element) is inducing damage on the final observed product.

Part Quality Contribution Indicator (PQCI):

The first algorithm uses a running log of parts produced by each machine. It checks the ratio of the quality of parts operated on by a particular machine versus the quality of parts that were not. It uses this comparison to determine an estimate of the average induced damage at that machine. For this study, the quality of the last ten parts to be operated on by each machine is kept and measured. One of the obvious shortcomings of this method is that if a machine does operate on a significant number of parts as compared to the total number produced, it will be slow to identify the problem machine. However, with even levels of part flow or if the problem element is not a high-risk machine, then this algorithm may be expected to perform suitably for many scenarios.

Estimated Part-Path Contribution Indicator (EPPCI):

The second algorithm investigated also makes use of a moving window of part quality to produce diagnostic indicators. However, this algorithm keeps running logs of each process path instead of one for each individual processing element. In scenarios with relatively few process paths, this may improve scalability. Conversely this could also hinder scalability in highly complex system configurations with a large number of paths.

By performing a simple calculation using the system configuration and path quality logs, this algorithm calculates the average product quality of each machine per path to identify both the best and the average part quality produced by each machine based on the paths that include it. This method attempts to circumvent the potential problem of having certain process links, instead of individual machines, be the source of induced errors. For this paper, isolation of links was strictly excluded as it is intended for future papers, but the setup of this algorithm begins that process. Regardless, this algorithm would also be expected to perform well in isolating individual problem machines and so is evaluated on such here.

3.3.2. Prediction Error Minimization Algorithms

There are a class of low entry barrier machine learning tools and algorithms that attempt to minimize some predictive function designed to describe the observed output of the system. The commonality of this class is the need for that descriptive function. In this case, a simple minimization between the observed product damage (1 – Quality) for each machine and the predicted total amount of added damage from the processing machines. Assuming each machine linearly adds a calculable amount of damage to the product during processing, this class of algorithms can estimate the average amount of that damage that is added by each machine.

$$EQ(1) \quad Error = abs(\,(1 - sum(IDam\_machs\_part)) - (1 - Q\_part)), \qquad for\ all\ parts$$

The selected predictive equation of error for a system, shown in Eq(1), is utilized for three different diagnostic algorithms. Based on simple Gradient Descent (GD), one on Genetic Algorithms (GA), and one on applied Neural Networks (NN). Regardless of the mechanical implementation, these tress routines are all essentially trying to minimize this error function by producing induced damage estimations (IDam) for each machine.

Gradient Descent (GD): The Gradient Descent based method utilizes a standard off-the-shelf product (Matlab's *fmincon*) to search the possibility space for the combination of each machining element between sensible limits set by the user. For simplicity in this test, the choice was made to limit the possible damage induced by each machine to be between 0 (no damage) and 1 (complete loss of product acceptance). While other implementations of this algorithm may find other bounds more suitable, these are simple enough to cover a broad range of cases and limit processing time to a functionally usable amount.

Genetic Algorithms (GA): Considered a generally more robust optimization method, genetic algorithms have been used extensively on optimization problems due to their broad applicability, ease of use, and global perspective [16]. Like the gradient descent method above, a genetic algorithm is applied to test case scenarios to minimize the objective function in Eq. (1). The optimization process also outputs each machine's contribution to part quality degradation and a prediction of the machines that have degraded. To ensure the feasibility of the predictions of machine degradations, the algorithm was iterated 5 times over each test case scenario and its predictions were averaged. The genetic algorithm implemented in Matlab's Global Optimization Toolbox was evaluated using its default options (i.e., population size = 50, maximum number of generations = 100).

Neural Networks (LSTM): The final prediction error minimization technique applied in this work is a Long Short Term Memory (LSTM) based neural network. These are a special form of recurrent neural networks, which as explained by [17], are better able to process time series information due to having storage potential creating an effective memory of local trends. Neural networks have become standard machine learning tools due to their ease of use, effectiveness as classifiers, and practicality for obtaining features out of a dataset. The LSTMs created for this work serve to provide not only preliminary insights on their applicability, but also to justify further testing on additional architectures and configurations in future studies.

The selected network configuration under evaluation has a sequence input layer, a series of two hidden LSTM layers (each followed by a 20 % dropout layer), one fully connected layer, and a final regression layer. The input vector to the LSTM network includes the part quality and a corresponding binary vector representing production path elements of that part. Creation and execution of the LSTM network was performed using Matlab's Deep Learning Toolbox. The specifics of hyperparameter selection and architecture development are beyond the scope of this work. The process used here to compare various algorithms could be equally suited for comparing various neural network architectures.

4. TEST RESULTS

The results of evaluating the algorithms discussed in Section 3.3 compare their ability to identify damage-inducing machines in different scenarios and begin to establish their expected performance for classes of scenarios. The five diagnostic algorithms are evaluated and compared by applying them to the test scenarios (see Table 2) developed for each of the two multistage manufacturing system configurations (see Table 1). That is 30 test case scenarios for system configuration #1, evenly divided into 3 different categories of test cases, and 48 test case scenarios for system configuration #2, evenly divided into 4 different categories.

Each algorithm is evaluated using metrics chosen to measure the algorithm's performance at diagnosing machine degradation. To relate to real world based effects, metrics that can be directly combined with negative scenario consequences were selected: the false negative rate (FNR), the false positive rate (FPR), and the combination of the two. The false positive rate is calculated as the number of good machines incorrectly predicted to be inducing damage divided by the total number of machines that are operating in the system (Type-I error). Likewise, the false negative rate is calculated as the number of machines inducing damage not isolated by the algorithm divided by the total number of machines that are indeed producing damaged units (Type-II error). Not identifying a machine as a source of damage incurs costs associated with manufacturing defective parts and goods. While incorrectly identifying a machine as a source of damage incurs costs from misspent maintenance, inspection, and production downtime.

Figures 3A and 3B summarize the evaluation results from applying each of the five diagnostic algorithms to the selected test scenarios. In system configuration #1 there are three classes of test case scenarios corresponding to the number of machines that are producing degraded products: one, two, or three machines. Each diagnostic algorithm, except the LSTM neural network, is applied over the ten scenarios in each category. The resulting FNR and FPR evaluations are averaged for each category. This is similarly done for system configuration #2, but with four different level categories of test case scenarios and twelve test cases in each category.

Due to the requirements for development of a neural network, the LSTM network testing was performed differently. Rather than evaluate its performance on each category of test scenarios individually, the LSTM was trained, tested, and evaluated on all the test case scenarios available for each system configuration (30 test cases for system #1 and 48 for system #2). The model randomly selected 67 % of each configuration's test case scenarios as training data and the rest of the scenarios as testing data. This ensured that the network had training of various conditions on each configuration, allowing it to produce predictions of both nominal and problem status of machines. This also means that the LSTM's evaluated FNR and FPR metrics in Figures 3A and 3B cannot be directly compared to the corresponding evaluations of the other algorithms, but it still points towards preliminary insights into the applicability of neural networks for machine failure predictions. The LSTM results are included here for completeness but will be readdressed in future work.

As structured in Section 3.3, the algorithms listed each produce relative values for problem likelihood. In order to translate this into a classification of 'problem', 'no problem', a series of thresholds must be established. A threshold of -0.1 likely induced damage is selected as a discriminator for evaluating PQCI and EPPCI. GD and GA have been framed to operate on a different scale which translates their cutoff to a threshold of 0.6 is selected as the discriminator for evaluating both GA and GD.

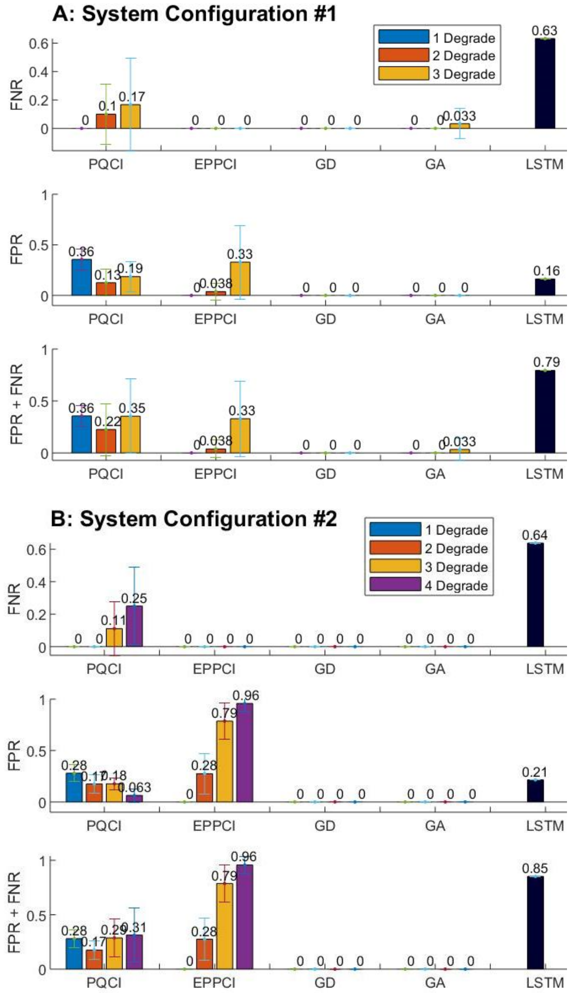


Figure 3. Evaluation results for applying the five diagnostic algorithms to predict machine degradation for three different test case scenario categories of: (A) system #1 and (B) system #2.

Like the GD and GA, the LSTM neural network also tries to predict which machines have contributed to part quality degradation. This means the threshold must also be between 0 and 1 to quantify the degree of a machine's contribution to part degradation. However, training the neural network on all the test case scenarios of each system configuration means that it tried learning from a wider range of antagonistic test case scenarios (rather than make a prediction from one test case at a time like the other four algorithms). This decreases the confidence in the predictions and so the discriminating threshold was lowered to 0.3.

At first glance, the results in Figures 3A and 3B show that the GD and GA algorithms provide superior performance to all the other algorithms. GD results in FNR and FPR values of 0, meaning it never identifies a failed machine as operational or an operational machine as degraded. This result is similar for GA, except a small false negative rate is produced when there are 3 machine degradations in system configuration #1. This supports a hypothesis that GD and GA will not make perfect predictions when there are too many machine failures in a production line. More testing is needed to see if this is indeed the case and to identify antagonistic test scenarios that may hamper the performance of GD and GA.

The PQCI and EPPCI algorithms exhibit trade-offs with each other. EPPCI produced no false negatives – no predictions of failed machines as operational - in any of the test case scenario levels for either system configuration. However, EPPCI produces a lot of false positives, especially for test cases where there are more machines that simultaneously degrade. It especially showed bad performance for multiple machine degradations in the second system configuration. On the other hand, PQCI produced false negatives that increased with more machine degradations, but it showed consistent levels of false positive rates that did not increase (drastically) with more machine degradations. The strength of one of these two algorithms over the other depends on the costs of having false negatives versus false positives in a manufacturing setting.

The predictive performance of both PQCI and EPPCI does not show to be as good as GD and GA, but their potential for scalability of more complex systems and to search for more indirect link based problems shows more promise. GD and GA attempt to search a full error space which grows exponentially with system complexity, while the evaluation space of the Q-learning space of PQCI and EPPCI grows linearly.

These insights are invaluable for creating antagonistic scenarios for further testing and evaluation of the algorithms. When identifying antagonistic test scenarios, an algorithm may have associated ranges of FNR and FPR that could help to further generalize the evaluation outcome. By properly combining the likelihood of these categories of test cases, their FNR and FPR values, an aggregated weighted evaluation of each diagnostic algorithm could be made to augment the existing test scenarios. This additional 'virtual testing' can save time and resources, while still valuably expand the coverage of the algorithm evaluation.

The LSTM neural network does provide many false negatives but produced false positive rates that are comparable to PQCI and EPCCI. A portion of the unfavorable results is due to training the LSTM on all test case scenarios for each system configuration. This points to the conclusion that learning on too many different antagonistic scenarios may not be as helpful for making machine failure predictions – there is a trade-off between the number of different antagonistic scenarios and the quality of the machine failure predictions. It will be helpful to see how the LSTM network performs when learning from and being applied to only one test case category level at a time. Furthermore, the LSTM network presented here is supposed to showcase preliminary work in the area, and more work will be done to fine-tune LSTM-related hyperparameters or evaluate other neural network structures.

5. CONCLUSIONS AND FUTURE WORK

This study showcases a preliminary understanding, applicability, and comparison of diagnostic algorithms (Section 3.3) when applied to predict machine failures in different simulated scenarios. The goal is to develop procedure recommendations for applying diagnostic monitoring algorithms across different scenarios, ranging from nominal expected scenarios to edge cases, faced by generic system configurations or particular configurations.

Further work needs to be done to extend the selection space of testing scenarios that are covered by the evaluation methodology, and to increase diagnostic algorithm development and evaluation. Two areas of particular interest right now include, 1) further development of risk assessment as an algorithm evaluation criteria, and 2) extending quality inspection to include mid-production process inspections. Algorithm evaluation can more precisely assess risk by incorporating costs and consequences associated with false negative and false positive machine degradation predictions, as well as relative likelihoods of different scenarios faced by the system configuration. This risk metric will represent more-practical effects and trade-offs of algorithms in relatable real-world measures beyond abstract classification metrics. In this preliminary study, the part quality inspection was conducted only at the end of the production path. This inspection data can be used to determine points in the production line that may require additional inspection or maintenance effort. These additional inspection points may provide diagnostic algorithms with more data collected throughout the production path and more context to make better predictions.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

[1] Lu, Biao & Zhou, Xiaojun. (2019). Quality and reliability oriented maintenance for multistage manufacturing systems subject to condition monitoring. Journal of Manufacturing Systems. 52. 76-85.

[2] Hao, Li & Bian, Linkan & Gebraeel, Nagi & Shi, Jianjun. (2016). Residual Life Prediction of Multistage Manufacturing Processes With Interaction Between Tool Wear and Product Quality Degradation. IEEE Transactions on Automation Science and Engineering. 14. 1-14. 10.1109/TASE.2015.2513208.

[3] Hu, S.J., Yoram Koren (1997). ”Stream-of-Variation Theory for Automotive Body Assembly”, CIRP Annals, Volume 46, Issue 1,1997, Pages 1-6.

[4] Ding, Yu & Ceglarek, Dariusz & Shi, Jianjun. (2000). Modeling and diagnosis of multistage manufacturing processes: Part I state space model.

[5] Djurdjanovic,D., J. Ni, (2001). "Linear state space modeling of dimensional machining errors", Trans. NAMRI/SME, vol. XXIX, pp. 541-548.

[6] Li, Yanting, & Fugee Tsung (2009) False Discovery Rate-Adjusted Charting Schemes for Multistage Process Monitoring and Fault Identification, Technometrics, 51:2, 186-205.

[7] Zeng, Li, & Shiyu Zhou (2007) Variability monitoring of multistage manufacturing processes using regression adjustment methods, IIE Transactions, 40:2, 109-121.

[8] Zhang, Min & Djurdjanovic, Dragan & Ni, Jun. (2007). Diagnosibility and sensitivity analysis for multi-station machining processes. International Journal of Machine Tools and Manufacture. 47. 646-657.

[9] Huang, Q., Zhou, S., and Shi, J. (2002). “Diagnosibility of Multi-Operational Machining Processes Through Variation Propagation Analysis,” Robotics and CIM Journal, 18, 233–239.

[10] Zhou, Shiyu, Yu Ding, Yong Chen & Jianjun Shi (2003) Diagnosability Study of Multistage Manufacturing Processes Based on Linear Mixed-Effects Models, Technometrics, 45:4, 312-325.

[11] Davari-Ardakani, H., & Lee, J. (2018). A Minimal-Sensing Framework for Monitoring Multistage Manufacturing Processes Using Product Quality Measurements.

[12] Liu, Y., Sun, R. and Jin, S. (2019), "A survey on data-driven process monitoring and diagnostic methods for variation reduction in multi-station assembly systems", Assembly Automation.

[13] Thoben, K. D., Wiesner, S., & Wuest, T. (2017). “Industrie 4.0” and smart manufacturing-a review of research issues and application examples. International Journal of Automation Technology, 11(1), 4-16.

[14] Koren, Yoram , S. Jack Hu, Thomas W. Weber, (1998) Impact of Manufacturing System Configuration on Performance, CIRP Annals, Volume 47, Issue 1.

[15] Drain, D. C. (1997). Handbook of experimental methods for process improvement. CRC Press.

[16] Goldberg, D. E. (1989). Genetic Algorithms for Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley.

[17] Cady, F. (2017). *The Data Science Handbook*. John Wiley & Sons.

Integrating a Network Simulator with the High Level Architecture for the Co-Simulation of Cyber-Physical Systems

Thomas Roth, Cuong Nguyen, and Martin Burns
Smart Grid and Cyber-Physical Systems Program Office
National Institute of Standards and Technology
Gaithersburg, MD 20899
thomas.roth@nist.gov, cuong.nguyen@nist.gov, martin.burns@nist.gov

Himanshu Neema
Institute for Software Integrated Systems
Vanderbilt University
Nashville, TN 37212
himanshu.neema@vanderbilt.edu

Keywords:
cyber-physical systems, high level architecture, network simulation, smart grid

ABSTRACT: *Cyber-physical systems (CPS) use logical computation informed by measurements of the environment to actuate changes on the physical world. These systems have significant impact on people, and must be designed for resilience against fault and attack. However, due to their large scale, assurance of CPS trustworthiness is better suited to modeling and simulation than deployment of a real system. This paper describes an approach to integrate a network simulator with the High Level Architecture (HLA) to investigate the effects of different network conditions on CPS performance. Using this approach, an HLA interaction class can be configured to use network simulation rather than the default reliable HLA delivery mechanism. A technique similar to regions defined in HLA data distribution management is used to allow each federate to receive the same interactions at different logical time steps, as simulated by the network simulator. This is implemented in a piece of reusable code shared by federates that sits between the runtime infrastructure (RTI) and application code. The implementation can be used to create a test harness around the federates that represent the operation of a CPS to validate its behavior under unreliable network conditions.*

1. Introduction

Cyber-physical systems (CPS) are smart systems that include engineered interacting networks of both physical and computational components [1]. These systems have a high degree of complexity at numerous spatial and temporal scales and need highly networked communications to integrate the computational and physical components. The smart grid is an example of a CPS that is defined as the integration of digital computing and communication technologies and services with the power-delivery infrastructure. The smart grid is often referred to as a system of systems that enables the bi-directional flow of both communication and power. Because the smart grid integrates information communication technology (ICT) with the electrical grid, network communication is one very important system component.

Grid operations are becoming more complex with the widespread deployment of distributed energy resources (DER) and distributed sensors that provide intelligence at the grid edge [2]. DER are comprised of many different types of resources such as solar photovoltaic (PV), wind, battery, and electric vehicle (EV). Some of these

Roth, Thomas; Nguyen, Cuong; Burns, Martin; Neema, Himanshu. "Integrating a Network Simulator with the High Level Architecture for the Co-Simulation of Cyber-Physical Systems." Paper presented at 2020 Simulation Innovation Workshop, Orlando, FL, US. February 10, 2020 - February 14, 2020.

resources such as PV and wind have volatility in energy production, and grid operators need robust communication capabilities to monitor and control them. Sensors distributed at the grid edge also have firm communication requirements to collect data, report, and fulfill their intended functions.

There are common standard-based communication protocols in use for grid operations. For substation automation, the two common protocols are IEC 61850 and Distributed Network Protocol 3 (DNP3). For DER communication, common protocols include IEEE 2030.5, DNP3, and SunSpec Modbus. Distributed sensors use protocols similar to both substation and DER. Beside these standardized protocols, there are manufacturer specific proprietary protocols. The choice of communication protocol for smart grid deployment depends on its performance capability, existing infrastructure, and the intended application. For example, if an operator plans to deploy an inverter for a PV installation, they need to consider what communication protocol their system can support to communicate with the inverter. If the operator plans to control DER, they will need a high-speed communication protocol for that application instead of a low bandwidth protocol that is only sufficient for monitoring alone. Other considerations may include whether the network connection between DER and the system is wired or wireless, and what cybersecurity mechanisms are required to protect the communication that still allow for the performance requirements.

Due to the complexity of the smart grid and its communication requirements, network simulation is essential and needs to represent the distributed nature of the evolving grid architecture. Sophisticated network simulation capabilities are needed to simulate the different communication protocols for the applications of interest to grid operators. Although such capabilities could be integrated into the implementation of grid simulators, it’s more intuitive to leverage the capabilities of existing network simulators. IEEE 1516-2010 High Level Architecture (HLA) is a standard for the co-simulation of distributed processes [3], such as the joint simulation of an electric grid and a network model. In HLA, the simulators that participate in a co-simulation are called federates, and the set of federates in the joint simulation are called a federation. The federates communicate and coordinate using software called the runtime infrastructure (RTI) that implements the common set of services described in the HLA federate interface specification [4]. An alternative to HLA is the Functional Mock-up Interface (FMI) standard often implemented by the developers of modeling tools [5]. Unlike HLA in which federates are independent processes in distributed system, FMI uses a master-slave architecture in which the master algorithm imports each simulator as a shared library and makes direct function calls into the slave code. The FMI standard for co-simulation prescribes the function definitions that each slave must implement to be interoperable with the master algorithm. This work uses HLA as its basis for co-simulation because it is more natural to consider a network simulator and a grid model as independent processes rather than sub-modules of one master program. In addition, networked co-simulation requires strict time management and distributed object management that are directly defined in HLA.

The remainder of this paper is organized as follows. Section 2 provides the motivation for the need of network simulation in the smart grid and lists the high level requirements that must be satisfied for network simulation to be meaningful in this context. Section 3 gives an overview of related work in the area, and Section 4 describes the specific approach to network simulation proposed by this work. The paper is then concluded in Section 5.

2. Motivation and Approach

For holistic system of systems evaluations, CPS require complex co-simulations including an integrated simulation of the cyber communication network as well as hardware- and human-in-the-loop. Owing to their use in critical system operations, the performance and trustworthiness of CPS must be evaluated under a variety of communication network modes which include the extreme cases of failures and attacks. In the smart grid, there are a variety of distributed sensors that are deployed at the system edge to provide situational awareness for monitoring and control. These sensors provide the condition at various points in the grid to detect any potential issues that could lead to system failure. One type of widely deployed sensor in the smart grid is the phasor measurement unit (PMU) or synchrophasor. PMUs provide voltage and current phasor and frequency measurements that are synchronized against a common time reference typically provided by global positioning system (GPS) [6]. Since these sensors provide time sensitive measurements, they need reliable communication to send the data to the grid control center. An attack on the timing infrastructure used by these devices could cause them to provide erroneous data to the operator that could lead to incorrect operating decisions such as unnecessary tripping of a line or not acting on a potential failure. Similarly, with the emergence of a plethora of innovative Internet of Things (IoT) devices for industrial control systems, edge computing, remote system monitoring and control, and home automation, it is equally critical to analyze the operational impacts of communication network failures for systems that incorporate IoT devices.

In order to analyze how communication network failures impact the operation of CPS and IoT, a careful consideration of the networked communication is necessary. In particular, analysis of a simple cyber attack might consider the impact of delaying the network packets, data corruption, replaying or reordering network packets, and packet loss. The HLA standard does not provide any direct means to support the co-simulation of these attack effects and supports only two options for delivering messages: receive order and timestamp order. In receive order, the messages are sent over UDP transport protocol and are delivered to the receiving federate with the best effort, without any explicit guarantee that a given message will be eventually delivered. In timestamp order, the messages are marked with a timestamp for delivery and are scheduled for delivery at that time to the receiving federate. A naïve approach may involve simply adding a delay to the timestamp of the delivered messages via HLA, but that does not realistically represent the behavior of unreliable message delivery. Even in cases in which a piece of manual code could be added to associated HLA federates' source code, this approach is highly inflexible and not representative of the flow of network packets in a real network. This is in contrast to using a communication network simulator that is integrated into the federation as a separate federate, where the networked communication between federates flows through the network simulator. The use of an integrated network simulator achieves not only faithful, high-fidelity network simulation, but also enables the realistic network characteristics such as unreliable message delivery against which the CPS and IoT systems can be evaluated.

The remainder of this section summarizes several desirable features for approaches to network simulation of CPS that are shown in Figure 1. The figure contains three federates, which consist of a federate implementation and the local RTI component (LRC) at each federate that implements the HLA message bus. Each federate has a unique representation in the simulated network model, depicted on the right. The remaining features in this figure are described in the subsections that follow.

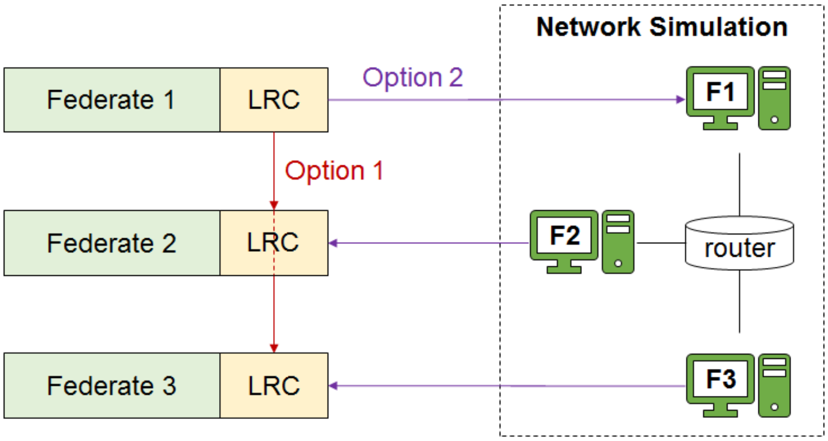


Figure 1: Approach to Network Simulation using the High Level Architecture (HLA)

2.1 The network model should contain nodes that represent a subset of the federates

At least two nodes from the network model should represent federates. The *Network Simulation* box from Figure 1 shows an example network model. A subset of the nodes in the network model indicated by labels F1 through F3 have a 1-to-1 correspondence to HLA federates. These federate nodes are connected through the simulated network topology. Although Figure 1 shows one network topology, the network model must be reconfigurable to allow the same federation to be executed with any number of different network configurations.

The network model does not need to define nodes to represent all federates in the federation. For instance, this federation could have a Federate 4 without the network model containing a corresponding node labeled F4. In this case, Federate 4 would not use the network simulation and all of its messages would use the default HLA provisions for object management. The network model also does not need to be fully connected. For instance, if node F1 cannot reach node F2 in the network topology, then none of the messages sent by Federate 1 using network simulation will be delivered to Federate 2.

2.2 The network simulator should be synchronized with the federation logical time

One responsibility of the network federate is to synchronize time progression of the simulated network with HLA federation logical time. The network federate is both time constrained and time regulating to operate in lock-step with HLA logical time. It also defines a function that maps a unit of HLA logical time to an exact number of seconds elapsed in the network simulation. This binding between the time representations of the federation and the network simulation ensures that a message is delivered to a federate only when the corresponding network packet is scheduled for delivery to that federate's node in the network simulation.

The optimal value for the logical step size of the network federate depends on the timing requirements of the federates using the simulated network. These timing requirements include concerns such as the smallest time interval between generation of network messages, and the shortest possible delivery time for messages sent from one federate to another. There is a trade-off between performance and simulation accuracy when choosing the logical step size. If the step size is too large, there will be delays in the delivery of messages to federates when a message arrives between time steps. If the step size is too small, the network simulator will synchronize more frequently with the federation which will lead to slower progression of logical time.

2.3 Network simulation should be configurable by both message type and sender

Even when a federate has a corresponding node in the network model, not all messages that originate from that federate are sent through the network simulation. A federate might want to coordinate with its peers or communicate with a federate that does not use the network simulation. For this reason, the use of network simulation is not configured per federate but rather per message that originates from a federate. This is shown in Figure 1 with two alternative paths for message flow listed as *Option 1* and *Option 2*. *Option 1* represents the normal HLA object management services where a federate can send and receive interactions and attribute updates using the RTI. *Option 2* is an alternative mode where specific messages are routed through a network simulator, rather than the usual set of HLA services.

In an ideal implementation, the LRC would perform this function of re-routing certain messages from the normal object management services into an alternative delivery mechanism based on the current network model. The RTI Initialization Data (RID) file could be modified to list the interactions and object classes that use network simulation. When the LRC received a message from the federate implementation, it would first check whether that specific message was configured for network simulation. If the message used simulation network, the LRC would send the message out-of-band to the network simulator. Otherwise, the LRC would continue to invoke the normal set of HLA object management services.

In this paper, the network simulator is a federate and *Option 2* is instead realized through re-encoding the message into a special interaction class that represents network packets. The network federate re-creates the original message once its corresponding network packet has propagated through the simulated network.

2.4 Federates should receive network simulated messages at different logical times

When the network federate receives a message from a LRC, it injects that message as one or more packets into the network simulation with the source of the packet set to the node representing the sending federate. If the network model is not based on multicast, then it is likely that one message will generate a unique packet for each federate node configured to receive that message that is reachable in the network model. All these packets will experience different delays, some may be dropped, and others might be modified through various forms of cyber-attacks. In the end, each federate node can receive a different packet, at different times, and perhaps with different content.

It is essential for cases such as network congestion and packet loss to break the reliable and uniform delivery of interactions and object classes guaranteed by HLA. Figure 1 shows an approach where this is implemented inside the RTI rather than the federates to reduce the amount of implementation required for each federate. However, the same effect could be achieved through implementation of a common library, shared by the federates, that sits between the LRC and the federate business logic.

2.5 Federate implementations should be agnostic to the presence of network simulation

Reusability is a desirable trait for federates developed for both CPS and IoT applications. Suppose a federate was developed to represent a PMU that reports time-synchronized voltage phasors measurements to some higher level application. This implementation could be useful for a number of different federation designs for different smart grid applications. Some of these applications might require realistic network delays to analyze the impacts of network congestion, some might require use of a specific communication protocol for hardware-in-the-loop testing, and some might just want to use a PMU with no network specific details. Despite differences in the interface on how the PMU is used, its basic implementation remains unchanged between these different applications.

While a typical HLA design flow might develop federations to achieve a specific purpose, in CPS and IoT applications, it is better to produce a federate like this PMU that can be composed into different scenarios. How this federate will be used is unknown at development time, and its implementation should support a broad range of scenarios without the need to develop additional code. For that reason, support for network simulation must be embedded into each federate as an option that can be enabled or disabled through configuration files. In addition, for different CPS and IoT federates to be interoperable, all the federates must implement their approach to optional network simulation using a consistent methodology.

3. Related Work

The integration of grid simulators and ICT into a co-simulation has over a decade of research. The first published approach in this area is the electric power and communication synchronizing simulator (EPOCHS) which uses HLA to integrate electromagnetic and electromechanical transient simulators with Network Simulator 2 (NS-2) [7]. Following EPOCHS, many co-simulation platforms were developed to integrate different grid simulators to different network simulators using different middleware [8]. The integrated co-simulation of power and ICT systems for real-time evaluation (INSPIRE) platform considers how to incorporate standard-based communication protocols into the co-simulation to support wide area monitoring, protection, and control (WAMPAC) applications [9]. The Hierarchical Engine for Large-scale Infrastructure Co-Simulation (HELICS) platform considers how to address scalability to handle grid scenarios that contain tens of thousands of independent agents [10].

The US National Institute of Standards and Technology (NIST) researched the effectiveness of different smart grid operating scenarios using the Framework for Network Co-Simulation (FNCS) developed by the Pacific Northwest National Laboratory that provides an integration of GridLAB-D, MATPOWER, and Network Simulator 3 (NS-3) [11]. The goal of this effort was to simulate a power grid segment that contained a substation and residential loads using different scenarios such as demand response and dynamic pricing. The work provided benchmarking for performance of the communication network under different operating conditions.

A follow-on work was on performance evaluation of DER and storage devices in terms of cost and impact on grid reliability [12]. This work was done by applying network traffic routing concepts to the routing of power in a grid segment with DER and storage devices. The premise for this work was that the resources are controllable, and the energy can be routed like network traffic management.

Additional research combined the smart grid operating scenarios (demand response and dynamic pricing) with the integration of DER in the grid [13]. This simulation work used a standard IEEE bus model with an integrated simulation platform that included GridMat, FNCS, GridLAB-D, and NS-3. The intent was to evaluate the performance of the grid with DER under different operating scenarios.

This paper attempts to address the feature from Section 2.4 on breaking the reliable delivery of HLA messages based on the results of network simulation. Other approaches largely limit their scope to adding message delays, and rarely consider the impact of packet loss or modification due to fault or cyber-attack.

4. Implementation Details

4.1 Universal CPS Environment for Federation (UCEF)

NIST has developed a software tool to expedite the development of federates and federations called the Universal CPS Environment for Federation (UCEF) [14]. UCEF is distributed as an Ubuntu virtual machine pre-configured with a suite of software useful in the development of different federate types. The latest 1.0.1 version of UCEF includes support for Java and C++ federates, and several grid simulators including GridLAB-D, TRNSYS, and EnergyPlus. The front end of UCEF is the Web-based Generic Modeling Environment (WebGME) developed at Vanderbilt University that provides a graphical web environment where users can model federations using simple building blocks. At the back end are JavaScript extensions to WebGME that perform code generation to transform the federate models into stub code for the different supported federate types. A core concept of UCEF is the separation of a federate implementation into two layers: a user layer that implements the intended function of the federate, and an infrastructure layer generated from WebGME that implements shared boiler plate code.

The UCEF infrastructure layer handles functions such as joining a federation, declaring publication and subscription interests, providing helper methods to send and receive interactions and object attributes, and other application independent utility functions. It also prescribes a basic federate lifecycle with hooks that an application developer can extend to customize the behavior of the federate at specific points in the HLA state machine, such as after the grant of an advance time request. The WebGME generated code closely resembles the structure of a Functional Mock-up Unit (FMU) as defined in the FMI standard for co-simulation.

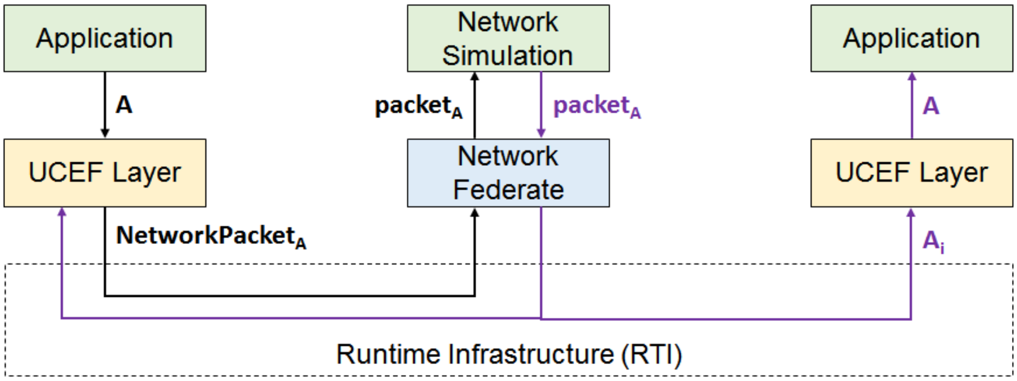


Figure 2: The UCEF Architecture for Network Simulation

Figure 2 shows how network simulation is implemented in UCEF to satisfy the requirements enumerated in Section 2. Rather than implement a new RTI, or modify an existing RTI, the logic related to network simulation was implemented in the UCEF infrastructure layer. In the figure, the user application on the left sends an interaction class A which has been flagged for network simulation. Before the UCEF layer passes this interaction to the RTI, it converts it into a special NetworkPacket interaction class to ensure it is routed to the network federate.

The network federate subscribes to the NetworkPacket interaction class, and encodes the received interaction instance as one or more packets in a format compatible with the network simulation. These packets are injected into the network simulation, where they propagate through the simulated network. When a packet arrives at its destination, it is delivered back to the network federate along with the name of the destination node from the network simulation. Then the network federate reconstructs the original interaction embedded inside the NetworkPacket and watermarks this new interaction with the destination node name. The network federate is implemented using a library called the UCEF Gateway [15], which enables it to create dynamic publications based on the current federation object model. As such, the network federate is not bound to a specific data model and can be used in any federation without code modification.

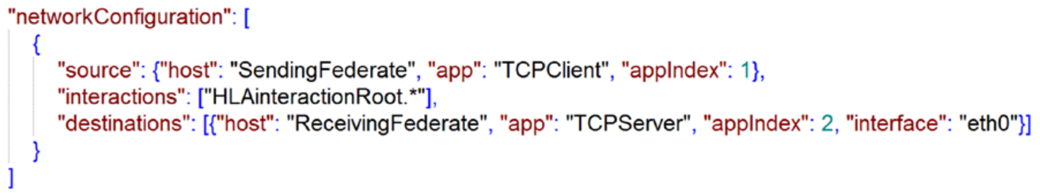
Because the network federate sends the watermarked interaction using the standard HLA services, all subscribed federates will receive it. However, each federate is configured to know the name of its representation in the network simulation, and the UCEF layer can use the watermark to check if an interaction was meant for its user application. In this manner, even though the left most federate receives a copy of its own interaction A_i¸ this interaction will be dropped at the UCEF layer before it reaches the user application.

One constraint of this approach is that all federates that use network simulation must have the same UCEF layer, which means they must be code generated from the UCEF virtual machine. The benefit of the approach is that the application never has to know about the network simulation and only must consider its native interaction class A. When network simulation is required, the infrastructure will take care of it. The remainder of this section describes how network simulation was embedded into the UCEF layer.

4.2 Network Configuration of UCEF Federates

There are three requirements to support network simulation in the UCEF layer: (1) the interaction classes that require network simulation must be specified in a configuration file, (2) the NetworkPacket interaction class must be defined, and (3) a filtering mechanism must be defined to ensure that interactions sent by the network federate are received only by the intended federates.

The same configuration file for network simulation is shared by all federates, including the network federate. This JavaScript Object Notation (JSON) file lists which interactions from which federates should use network simulation and how those interactions should be routed through the network simulation. Figure 3 shows an example instance of this configuration file. This JSON configuration was designed for use with the OMNeT++ network simulator using its INET Framework.



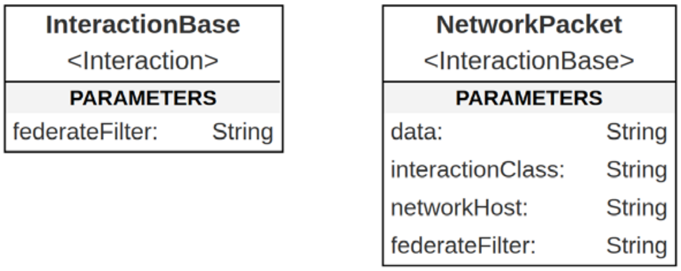
```
"networkConfiguration": [
    {
        "source": {"host": "SendingFederate", "app": "TCPClient", "appIndex": 1},
        "interactions": ["HLAinteractionRoot.*"],
        "destinations": [{"host": "ReceivingFederate", "app": "TCPServer", "appIndex": 2, "interface": "eth0"}]
    }
]
```

Figure 3: Example JSON for Network Configuration

The network configuration is a list of network rules. Each network rule defines the list of *interactions* for a given *source* federate that are configured to use network simulation. These interactions are injected into the network model at the *source* node and routed to each of the listed *destinations*. The same *source* federate can appear in multiple network rules for the case when different interactions from the same source have different destinations. This implementation assumes for simplicity that the federate name is identical to the host name of its equivalent network node. Under this assumption, the *host* fields are both the federate name and the network node name.

A network node in the OMNeT++ INET Framework contains submodules for different network applications. For example, a node could define a submodule for a Representational State Transfer (REST) server running on localhost:8080. These network applications are identified using an application name (*app*) and an application or submodule index (*appIndex*). The destination applications are bound to a specific network interface (*interface*) to handle cases where a node has more than one available network interface.

Figure 4 shows the WebGME representations of the two interaction classes *HLAinteractionRoot.InteractionBase* and *HLAinteractionRoot.InteractionBase.NetworkPacket*.



InteractionBase
<Interaction>

PARAMETERS	
federateFilter:	String

NetworkPacket
<InteractionBase>

PARAMETERS	
data:	String
interactionClass:	String
networkHost:	String
federateFilter:	String

Figure 4: Example Object Model for Network Simulation

The UCEF layer encodes all interactions it sends to the network federate using the NetworkPacket interaction class. Besides the *federateFilter* parameter that will be discussed later, this interaction class has three parameters. The *interactionClass* and *data* parameters are used to embed the original interaction from the user application into the NetworkPacket. When an instance of interaction class A is converted into a NetworkPacket, the *interactionClass* parameter would be the fully qualified class path of A and the *data* parameter would be the serialized parameters of A. These fields are used by the network federate to reconstruct the original interaction after the packet propagates through the network simulation. The final parameter, *networkHost*, is set to the unique identifier of the sending federate. This allows the network federate to inject the packet into the network simulation at the correct node.

When the UCEF layer receives an interaction, it checks if that interaction was sent by the network federate. If the interaction was sent by the network federate, then it is possible that it was only intended for receipt by a single federate. The network federate uses the *federateFilter* parameter to specify this destination. If the *federateFilter* is empty, then the UCEF layer processes the interaction as normal. Otherwise, the UCEF layer discards the interaction unless the *federateFilter* is string equivalent to the federate's own unique identifier.

4.3 Implementation of the UCEF Layer

Figure 5 shows a flowchart for how interaction classes are handled from both the sending and the receiving federates. The first decision box at the top of the figure, whether to use network simulation, is determined by the content of the JSON file from Figure 3. The second decision box at the bottom of the figure, whether the received interaction should be delivered to this specific federate, performs the string comparison between the *federateFilter* parameter from Figure 4 and the federate name of the receiving federate. If the federate filter parameter is set but not equivalent to the receiving federate name, then the packet is dropped and not delivered to the user application.

This filtering mechanism allows different federates to receive the same interaction at different logical times, dependent on the results of network simulation. An equivalent implementation could have been achieved using regions from the HLA data distribution management (DMM) services instead of as a parameter of a base interaction type. However, as the configuration management of regions can be quite cumbersome, this simple filtering mechanism was implemented at the UCEF layer instead.

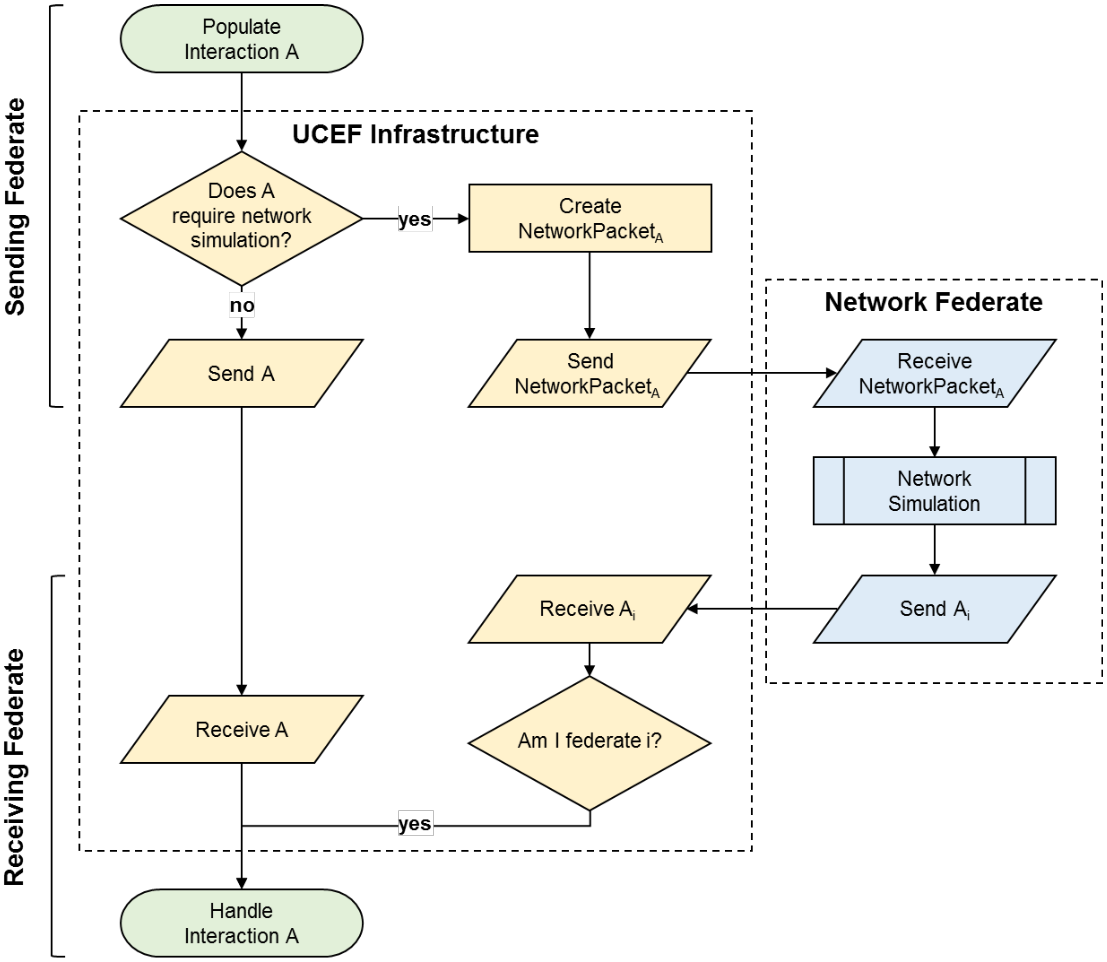
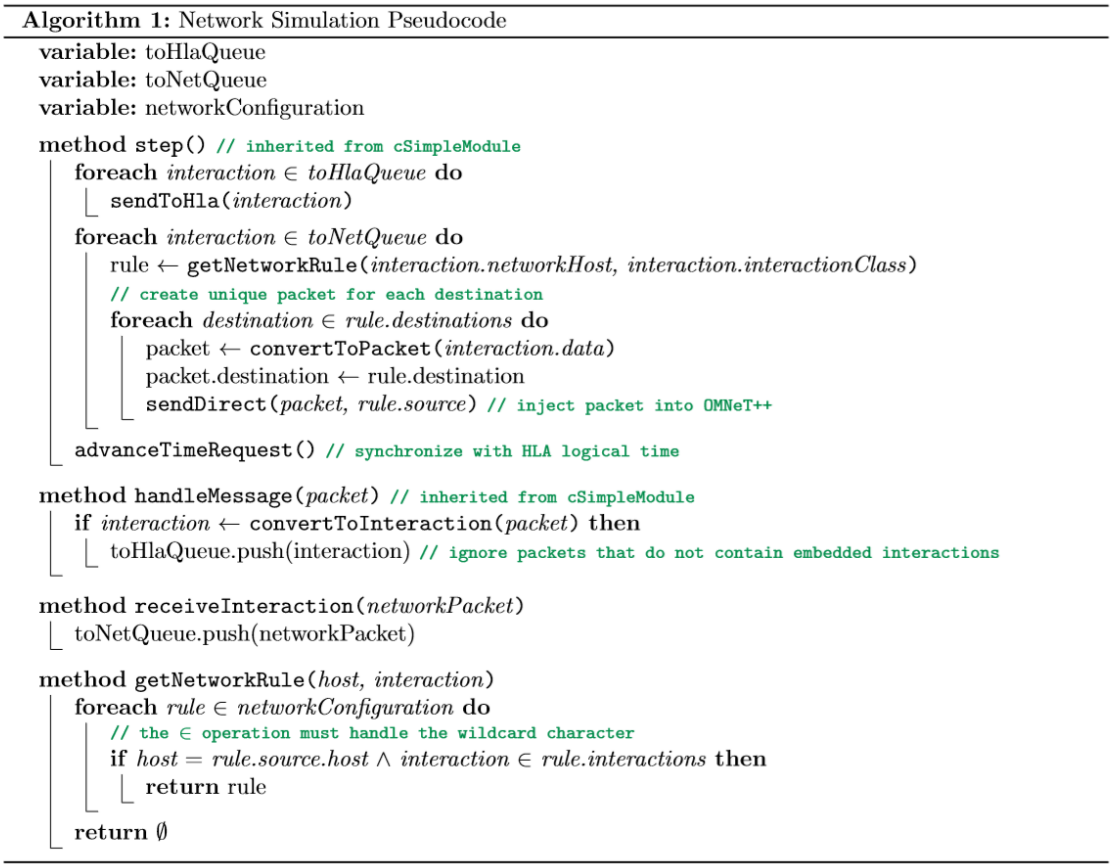


Figure 5: Flow of Interactions through the UCEF Layer

4.4 Implementation of the Network Federate

Algorithm 1 shows pseudocode for the network simulation from Figure 5. This algorithm uses the OMNeT++ network simulator, and extends the OMNeT++ *cSimpleModule* class which defines the *step* and *handleMessage* methods. The *step* method executes each HLA logical time step and checks for either packets from the network simulator or interactions from the federation. When an interaction is received, the *step* method parses the JSON network configuration to create packets in the network simulation for each configured destination. These packets propagate through the simulated network until they arrive at their destination, causing OMNeT++ to invoke the *handleMessage* method. In *handleMessage*, a customized interaction is created for the node that received the packet. The current implementation of time synchronization maps a unit of HLA logical time to a second of network simulation time. The *advanceTimeRequest* method could be replaced with an alternative implementation that uses a scaling function to make each logical time step some configurable multiple of seconds.

Roth, Thomas; Nguyen, Cuong; Burns, Martin; Neema, Himanshu. "Integrating a Network Simulator with the High Level Architecture for the Co-Simulation of Cyber-Physical Systems." Paper presented at 2020 Simulation Innovation Workshop, Orlando, FL, US. February 10, 2020 - February 14, 2020.



Algorithm 1: Network Simulation Pseudocode

```
variable: toHlaQueue
variable: toNetQueue
variable: networkConfiguration

method step() // inherited from cSimpleModule
    foreach $interaction \in toHlaQueue$ do
        sendToHla($interaction$)
    foreach $interaction \in toNetQueue$ do
        rule ← getNetworkRule($interaction.networkHost$, $interaction.interactionClass$)
        // create unique packet for each destination
        foreach $destination \in rule.destinations$ do
            packet ← convertToPacket($interaction.data$)
            packet.destination ← rule.destination
            sendDirect($packet$, $rule.source$) // inject packet into OMNeT++
    advanceTimeRequest() // synchronize with HLA logical time

method handleMessage($packet$) // inherited from cSimpleModule
    if $interaction$ ← convertToInteraction($packet$) then
        toHlaQueue.push(interaction) // ignore packets that do not contain embedded interactions

method receiveInteraction($networkPacket$)
    toNetQueue.push(networkPacket)

method getNetworkRule($host$, $interaction$)
    foreach $rule \in networkConfiguration$ do
        // the ∈ operation must handle the wildcard character
        if $host = rule.source.host \wedge interaction \in rule.interactions$ then
            return rule
    return $\emptyset$
```

5. Conclusion

This paper proposed an approach to incorporate network simulation into the High Level Architecture (HLA). Whether the proposed approach is HLA compliant depends on the interpretation of the rule that *"During a federation execution, all exchange of FOM data among federates shall occur via the RTI"* [4]. From the federate perspective, this rule is upheld because the NetworkPacket interaction is transmitted to the federation via the RTI. However, from the user application perspective, the interaction that the user wants to send is automatically converted into a different interaction class and the RTI is never used to transmit the data in its intended format. This approach was chosen because co-simulation of CPS requires network simulations that support the concepts of message delay and packet drop, and integrating the semantics of network simulation into each individual federate - while feasible - is far too burdensome.

The next step is to complete the implementation of this approach in UCEF. An implementation that addresses the need for packet loss and modification was developed by Vanderbilt University for their Command and Control Wind Tunnel (C2WT) platform [16][17]. However, it requires source code modifications when changes are made to the network configuration. Another implementation was released by Calytrix for their open-source Portico RTI [18]. However, it requires the use of multiple network federates which leads to poor scalability for large federation sizes, and the approach may be incompatible with the popular OMNeT++ INET Framework. Future work will merge these implementations to produce one network federate configurable using JSON that is compatible with the OMNeT++ INET Framework. This future work will aim to improve the time synchronization strategy to be more flexible than a 1-to-1 equivalence between HLA logical time and the network simulator time.

Roth, Thomas; Nguyen, Cuong; Burns, Martin; Neema, Himanshu. "Integrating a Network Simulator with the High Level Architecture for the Co-Simulation of Cyber-Physical Systems." Paper presented at 2020 Simulation Innovation Workshop, Orlando, FL, US. February 10, 2020 - February 14, 2020.

6. Acknowledgement

Portions of this publication and research effort are made possible through the help and support of NIST via cooperative agreement 70NANB19H100. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

7. References

[1] Griffor, E. R., Greer, C., Wollman, D. A., & Burns, M. J. (2017). *Framework for cyber-physical systems: Volume 1, overview* (NIST-SP-1500-201). doi: 10.6028/NIST.SP.1500-201

[2] Greer, C., Wollman, D., Prochaska, D., Boynton, P., Mazer, J., Nguyen, C., FitzPatrick, G., Nelson, T., Koepke, G., Hefner Jr., A., Pillitteri, V., Brewer, T., Golmie, N., Su, D., Eustis, A., Holmberg, D., & Bushby, S. (2014). *NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 3.0*. (NIST SP-1108r3) doi: 10.6028/NIST.SP.1108r3

[3] Institute of Electrical and Electronics Engineers. (2010). *IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-- Framework and Rules* (IEEE Std 1516-2010) doi: 10.1109/IEEESTD.2010.5553440

[4] Institute of Electrical and Electronics Engineers. (2010). *IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-- Federate Interface Specification* (IEEE Std. 1516.1-2010) doi: 10.1109/IEEESTD.2010.5557728

[5] *Functional Mock-up Interface for Model Exchange and Co-Simulation 2.0* (2014, July). Retrieved November 27, 2019 from http://fmi-standard.org

[6] Terzija, V. (2011). Wide-Area Monitoring, Protection, and Control of Future Electric Power Networks. *Proceedings of the IEEE*, 99(1), 80-93. doi: 10.1109/JPROC.2010.2060450

[7] Hopkinson, K., Wang, X., Giovanini, R., Thorp, J., Birman, K., & Coury, D. (2006). EPOCHS: a platform for agent-based electric power and communication simulation built from commercial off-the-shelf components. *IEEE Transactions on Power Systems*, *21*(2), 548-558. doi: 10.1109/TPWRS.2006.873129

[8] IEEE Task Force on Interfacing Techniques for Simulation Tools (2016). Interfacing Power System and ICT Simulators: Challenges, State-of-the-Art, and Case Studies. *IEEE Transactions on Smart Grid, 9*(1), 14-24. doi: 10.1109/TSG.2016.2542824

[9] Georg, H., Müller, S. C., Rehtanz, C., & Wietfeld, C. (2014). Analyzing cyber-physical energy systems: The INSPIRE cosimulation of power and ICT systems using HLA. *IEEE Transactions on Industrial Informatics*, *10*(4), 2364-2373. doi: 10.1109/TII.2014.2332097

[10] Palmintier, B., Krishnamurthy, D., Top, P., Smith, S., Daily, J., & Fuller, J. (2017, April). Design of the HELICS high-performance transmission-distribution-communication-market co-simulation framework. In *2017 Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES)* (pp. 1-6). IEEE. doi: 10.1109/MSCPES.2017.8064542

[11] Moulema, P., Yu, W., Griffith, D., & Golmie, N. (2015). On Effectiveness of Smart Grid Applications Using Co-Simulation. *24th International Conference on Computer Communication and Networks (ICCCN)*. doi: 10.1109/ICCCN.2015.7288438

[12] Xu, G., Yu, W., Griffith, D., Golmie, N., & Moulema, P. (2016). Towards Integrating Distributed Energy Resources and Storage Devices in Smart Grid. *IEEE Internet of Things Journal*. 4(1): pp 192-204. doi: 10.1109/JIOT.2016.2640563

[13] Mallapuram, S., Yu, W., Moulema, P., Griffith, D., Golmie, N., & Liang, F. (2017). An Integrated Simulation Study on Reliable and Effective Distributed Energy Resources in Smart Grid. *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. pp 140-145. doi: 10.1145/3129676.3129684

[14] Burns, M., Roth, T., Griffor, E., Boynton, P., Sztipanovits, J., & Neema, H. (2018). Universal CPS Environment for Federation (UCEF). In *2018 Winter Simulation Innovation Workshop*.

[15] Roth, T., & Burns, M. (2018). A gateway to easily integrate simulation platforms for co-simulation of cyber-physical systems. In *2018 Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES)* (pp. 1-6). IEEE. doi: 10.1109/MSCPES.2018.8405394

[16] Hemingway, G., Neema, H., Nine, H., Sztipanovits, J., & Karsai, G. (2012). Rapid synthesis of high-level architecture-based heterogeneous simulation: a model-based integration approach. *Simulation, 88*(2), 217-232. doi: 10.1177/0037549711401950

[17] Neema, H. (2018). *Large-Scale Integration of Heterogeneous Simulations* (Doctoral dissertation, Vanderbilt University). Retrieved from https://www.isis.vanderbilt.edu/node/4925

[18] *Federate Base* (2019, July). Retrieved November 27, 2019 from https://github.com/openlvc/federate-base

Author Biographies

THOMAS ROTH leads development of the technology behind the cyber-physical systems testbed at the National Institute of Standards and Technology as a member of its Smart Grid and Cyber-Physical Systems program office. His research interests are in formal methods for the composition of cyber-physical systems, and the detection of compromised cyber-physical devices through comparison of their reported behavior against the constraints of the physical system.

CUONG NGUYEN leads the Smart Grid Testing and Certification Project in the Smart Grid and Cyber-Physical Systems Program Office of the Engineering Laboratory at the National Institute of Standards and Technology. He works with industry to support standards-based interoperability test programs to help accelerate smart grid deployments. Cuong is the chair of the Smart Electric Power Alliance (SEPA) Testing and Certification Working Group (TCWG). Cuong coordinates international outreach efforts through bilateral and multilateral engagements.

MARTIN BURNS is the Associate Director for the CPS/IoT Testbed in the Smart Grid and Cyber-Physical Systems Program Office at NIST. With his background in IEC and ANSI standards development for semantic models and data exchange, he has facilitated the development of the underlying Green Button technologies which define energy usage information and APIs in the Smart Grid in the US and internationally. He co-chairs the data interoperability working group for the NIST led Framework for Cyber-Physical Systems (CPS) and is a key contributor to NISTs architecture for UCEF-federated testbeds for investigating the behaviors of CPS/IoT.

HIMANSHU NEEMA is a Research Assistant Professor of Computer Science at Vanderbilt University. He holds a M.S. and Ph.D. in Computer Science from Vanderbilt University. Dr. Neema researches in the general area of model-based design and modeling and simulation of Cyber-Physical Systems and their integrated simulation with hardware- and humans- in the loop. His research interests include: Modeling & Simulation, Model-Integrated Computing, Distributed Simulations, Artificial Intelligence, Constraint Programming, Planning & Scheduling, Smart-Grids, Transactive Energy, Service-Oriented Architectures (SOAs), Semantic Web, and Automated Document Analysis & Classification. Dr. Neema has 20 years of experience in research and development of software applications covering the above areas and has co-authored ~50 publications.

17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020
Paper N° C002026
Registration Code: S-A02125

IDENTIFICATION OF GEOMETRIC NONLINEARITY FORMULATIONS AND THE INFLUENCES ON STRUCTURAL DYNAMIC RESPONSES

K.K.F. Wong[(1)]

[(1)] *Research Structural Engineer, National Institute of Standards and Technology, USA, kfwong@nist.gov*

Abstract

Both ASCE/SEI 7-16 [1] and ASCE/SEI 41-17 [2] design standards require that P-Δ effects be included in the nonlinear dynamic analysis procedure for designing new and existing structures. However, software packages that perform nonlinear dynamic analysis typically only allow users to either 'turn on' or 'turn off' the effect of geometric nonlinearity without informing users what type of nonlinear effect is being included, such as whether both large P-Δ and small P-δ effects are included. Coupling this geometric nonlinearity effect with material nonlinearity of the structure can lead to significant differences in the predicted responses, especially when the structure is at near-collapse. Therefore, study on how geometric nonlinearity can affect structural performance is important to performance-based seismic engineering.

The challenge of how each software package handles the coupling between geometric and material nonlinearities often comes down to their interaction. The traditional method of structural analysis with material nonlinearity uses changing stiffness to quantify the stiffness reduction after yielding has occurred in the structure, which requires reformulation of the global stiffness matrix before solving for the displacement response. The difficulty comes in when the axial force reduces the stiffness of the member and at the same time when tangent stiffness is used after a nonlinear component has yielded. Will this axial force further reduce the tangent stiffness, and by how much? On the other hand, the use of tangent stiffness is typically derived based only on material nonlinearity with no consideration of geometric nonlinearity. At the same time, the use of geometric stiffness was derived based only on geometric nonlinearity with no consideration of material nonlinearity. Therefore, one needs to go back to the original derivation with the basic principles to answer this question.

In this paper, a detailed formulation of the nonlinear stiffness matrices taking into consideration both geometric and material nonlinearities is derived for a column member. It is followed by a study on how different software packages implement different geometric nonlinearity formulations and identify whether such formulation include both large P-Δ and small P-δ effects in the analysis. Finally, examples are presented to demonstrate how different formulations of geometric nonlinearity affect the nonlinear responses of framed structures.

Keywords: Geometric nonlinearity; Material nonlinearity; Large P-Δ effect; Small P-δ effect; Displacement

Wong, Kevin K F. "Identification of Geometric Nonlinearity Formulations and the Influences on Structural Dynamic Responses." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Buildings constructed in seismic regions are vulnerable to strong ground shaking and thus are designed to sustain damage caused by material nonlinearity and remain stable by overcoming geometric nonlinearity. Nonlinear dynamic analysis with both geometric and material nonlinearities is currently the most accurate method of estimating the response to seismic events. For this reason, both ASCE/SEI 7-16 [1] and ASCE/SEI 41-17 [2] require that geometric nonlinearity be included in the nonlinear dynamic analysis procedure. However, software packages provide a geometric nonlinearity "button" but do not explicitly state the specific geometric nonlinearity formulation implemented in these packages. For example, some software packages may implement the full geometric stiffness formulation that includes both large *P*-Δ and small *P*-δ effects, while other software packages may implement the simpler *P*-Δ stiffness formulation that accounts for only *P*-Δ effect.

This research investigates how different software packages implement geometric nonlinearity affect the nonlinear dynamic response calculations. To accomplish this work, a detailed formulation of the nonlinear stiffness matrices taking into consideration both geometric and material nonlinearities is first derived for a column member. It is followed by a study on how different software packages implement different geometric nonlinearity formulations and identify whether such formulation include both large *P*-Δ and small *P*-δ effects in the analysis. Finally, examples are presented to demonstrate how different geometric nonlinearity formulations affect the nonlinear responses of framed structures.

2. Derivations of Stiffness Matrix with Both Geometric and Material Nonlinearities

The original beam theory with geometric nonlinearity [3-4] was first developed for elastic columns in the 1960's without consideration of any material nonlinearity. But its use is limited because of its complexity in the closed-form solution as compared to those formulations based on either the *P*-Δ stiffness approach [5] or the geometric stiffness approach [6]. However, to identify the geometric nonlinearity used in software packages, it is often important to go back to the fundamental principle and understand how each geometric nonlinearity formulation is developed. Therefore, the original theory that included geometric nonlinearity is rederived here and extended to include material nonlinearity. This is done by deriving the stiffness matrix for a column member with plastic hinges at both ends and subjected to a compressive force.

Four degrees of freedom (DOFs) and two plastic hinge locations (PHLs) are used to describe the movements at the two ends of a column member in a moment-resisting frame. These movements at the two ends include lateral displacement ($v(0)$ and $v(L)$), rotation ($v'(0)$ and $v'(L)$), and plastic rotations at the two plastic hinges (θ''_a and θ''_b). To compute the member stiffness matrix $\mathbf{k}_i$, where *i* denotes the i^{th} member in the frame, each of these 4 DOFs and 2 PHLs is displaced independently by one unit as shown in Fig.1 while subjected to an axial compressive load *P*. Here, V_{1k}, M_{1k}, V_{2k}, and M_{2k} represent the required shear forces and moments at the two ends of the member to cause the deformation in the prescribed pattern, where $k = 1,\ldots,6$ represents the six cases of unit displacement patterns of the member's movements, and M_{ak} and M_{bk} represent the moment at plastic hinges '*a*' and '*b*', respectively, due to the prescribed pattern. Note that the '1' end coincides with plastic hinge '*a*' and the '2' end coincides with plastic hinge '*b*' in Fig.1.

Using the classical Bernoulli-Euler beam theory with homogeneous and isotropic material properties, where the moment is proportional to the curvature and plane sections are assumed to remain plane based on small displacements, the governing equilibrium equation describing the deflected shape of the member is

$$(EIv'')'' + Pv'' = 0 \tag{1}$$

where *E* is the elastic modulus, *I* is the moment of inertia, *v* is the lateral deflection, *P* is the axial compressive force on the member, and each prime represents taking derivatives of the corresponding variable with respect to the *x*-direction of the member. By assuming *EI* is constant along the member, the solution to the fourth-order ordinary differential equation given in Eq. (1) becomes:

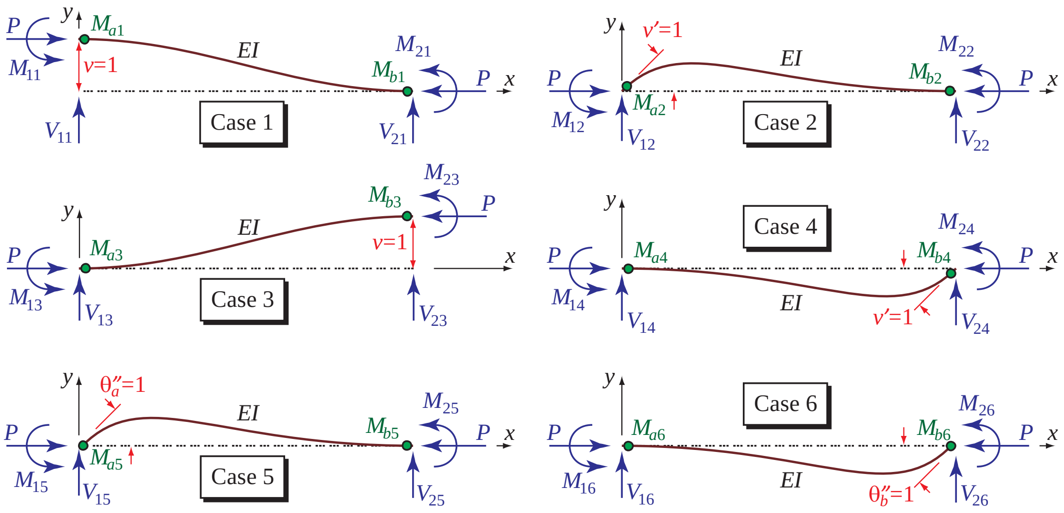


Fig. 1 – Six cases of unit displacement patterns and corresponding fixed-end forces and hinge moments

$$v = A\sin kx + B\cos kx + Cx + D \tag{2}$$

where $k^2 = P/EI$ and A, B, C, and D are constants to be determined by imposing different boundary conditions. Let $\lambda = kL$ to simplify the derivations, where L is the length of the member. The following cases of boundary conditions (starting with Case 4) are now considered.

2.1 Case 4

For Case 4 as shown in Fig.1, imposing the boundary conditions $v(0)=0$, $v'(0)=0$, $v(L)=0$, $v'(L)=1$, and $\theta''_a = \theta''_b = 0$ on Eq. (2) gives

$$v(0)=0: \qquad B + D = 0 \tag{3a}$$

$$v'(0)=0: \qquad kA + C = 0 \tag{3b}$$

$$v(L)=0: \qquad A\sin\lambda + B\cos\lambda + CL + D = 0 \tag{3c}$$

$$v'(L)=1: \qquad kA\cos\lambda - kB\sin\lambda + C = 1 \tag{3d}$$

Solving simultaneously for the constants in Eq. (3) gives

$$A = \frac{L(1-\cos\lambda)}{\lambda(\lambda\sin\lambda + 2\cos\lambda - 2)} \ , \quad B = \frac{L(\sin\lambda - \lambda)}{\lambda(\lambda\sin\lambda + 2\cos\lambda - 2)} \ , \quad C = -kA \ , \quad D = -B \tag{4}$$

Therefore, Eq. (2) along with the constants in Eq. (4) gives the deflected shape for Case 4. The shears (i.e., V_{14} and V_{24}) and moments (i.e., M_{14} and M_{24}) at the two ends of the member (see Fig.1) are then evaluated using the classical Bernoulli-Euler beam theory formula:

$$M(x) = EIv'' \quad , \quad V(x) = EIv''' + Pv' \tag{5}$$

Now taking derivatives of Eq. (2) and substituting the results into Eq. (5) while using the constants calculated in Eq. (4), the shears and moments at the two ends of the member are calculated as:

$$M_{14} = -EIv''(0) = EIk^2B = \hat{s}\hat{c}EI/L \tag{6a}$$

$$V_{14} = EIv'''(0) + Pv'(0) = -EIk^3A + P\times 0 = \bar{s}EI/L^2 \tag{6b}$$

$$M_{24} = EIv''(L) = -EIk^2(A\sin\lambda + B\cos\lambda) = \hat{s}EI/L \tag{6c}$$

$$V_{24} = -EIv'''(L) - Pv'(L) = EIk^3\left(A\cos\lambda - B\sin\lambda\right) - P\times 1 = -\bar{s}EI/L^2 \tag{6d}$$

where $\hat{s}$, $\hat{c}$, and $\bar{s}$ are the first three stability coefficients computed by the formula

$$\hat{s} = \frac{\lambda\left(\sin\lambda - \lambda\cos\lambda\right)}{2-2\cos\lambda-\lambda\sin\lambda} \quad , \quad \hat{c} = \frac{\lambda - \sin\lambda}{\sin\lambda - \lambda\cos\lambda} \quad , \quad \bar{s} = \hat{s} + \hat{s}\hat{c} = \frac{\lambda^2\left(1-\cos\lambda\right)}{2-2\cos\lambda-\lambda\sin\lambda} \tag{7}$$

In addition, the moments at the two PHLs can be evaluated by recognizing that these moments must equal to the end moments by equilibrium, i.e., $M_{a4} = M_{14}$ and $M_{b4} = M_{24}$. Therefore,

$$M_{a4} = M_{14} = \hat{s}\hat{c}EI/L \quad , \quad M_{b4} = M_{24} = \hat{s}EI/L \tag{8}$$

2.2 Case 3

For Case 3 as shown in Fig.1, imposing the boundary conditions $v(0)=0$, $v'(0)=0$, $v(L)=1$, $v'(L)=0$, and $\theta''_a = \theta''_b = 0$ on Eq. (2) gives

$$v(0)=0: \qquad B + D = 0 \tag{9a}$$
$$v'(0)=0: \qquad kA + C = 0 \tag{9b}$$
$$v(L)=1: \qquad A\sin\lambda + B\cos\lambda + CL + D = 1 \tag{9c}$$
$$v'(L)=0: \qquad kA\cos\lambda - kB\sin\lambda + C = 0 \tag{9d}$$

Solving simultaneously for the constants in Eq. (9) gives

$$A = -\frac{\sin\lambda}{\lambda\sin\lambda + 2\cos\lambda - 2} \quad , \quad B = \frac{1-\cos\lambda}{\lambda\sin\lambda + 2\cos\lambda - 2} \quad , \quad C = -kA \quad , \quad D = -B \tag{10}$$

These constants in Eq. (10) are used to give the deflected shape in Eq. (2) for Case 3. Now substituting Eq. (2) into Eq. (5) and using the constants calculated in Eq. (10), the shears and moments at the two ends of the member are calculated as:

$$M_{13} = -EIv''(0) = EIk^2B = -\bar{s}EI/L^2 \tag{11a}$$
$$V_{13} = EIv'''(0) + Pv'(0) = -EIk^3A + P\times 0 = -s'EI/L^3 \tag{11b}$$
$$M_{23} = EIv''(L) = -EIk^2\left(A\sin\lambda + B\cos\lambda\right) = -\bar{s}EI/L^2 \tag{11c}$$
$$V_{23} = -EIv'''(L) - Pv'(L) = EIk^3\left(A\cos\lambda - B\sin\lambda\right) - P\times 0 = s'EI/L^3 \tag{11d}$$

where s' is the fourth and final stability coefficient given by the formula

$$s' = 2\bar{s} - \lambda^2 = \frac{\lambda^3\sin\lambda}{2-2\cos\lambda-\lambda\sin\lambda} \tag{12}$$

In addition, the moments at the two PHLs are evaluated by equilibrium as

$$M_{a3} = M_{13} = -\bar{s}EI/L^2 \quad , \quad M_{b3} = M_{23} = -\bar{s}EI/L^2 \tag{13}$$

2.3 Case 2

For Case 2 as shown in Fig.1, by imposing the boundary conditions $v(0)=0$, $v'(0)=1$, $v(L)=0$, $v'(L)=0$, and $\theta''_a = \theta''_b = 0$ on Eq. (2), solution can be obtained via the same procedure presented above while solving for a different set of constants. On the other hand, a more direct solution can be obtained by recognizing that Case 2 is exactly the same as ‘rotating’ Case 4 by 180°. Doing so, the solution becomes

$$V_{12} = -V_{24} = \bar{s}EI/L^2 \quad , \quad V_{22} = -V_{14} = -\bar{s}EI/L^2 \tag{14a}$$
$$M_{12} = M_{24} = \hat{s}EI/L \quad , \quad M_{22} = M_{14} = \hat{s}\hat{c}EI/L \tag{14b}$$
$$M_{a2} = M_{b4} = \hat{s}EI/L \quad , \quad M_{b2} = M_{a4} = \hat{s}\hat{c}EI/L \tag{14c}$$

2.4 Case 1

For Case 1 as shown in Fig.1, by imposing the boundary conditions $v(0)=1$, $v'(0)=0$, $v(L)=0$, $v'(L)=0$, and $\theta''_a=\theta''_b=0$ on Eq. (2), solution can be obtained via the same procedure presented above while solving for a different set of constants. On the other hand, a more direct solution can be obtained by recognizing that Case 1 is exactly the same as 'flipping' Case 3 by 180°. Doing so, the solution becomes

$$V_{11}=V_{33}=s'EI/L^3 \quad , \quad V_{21}=V_{13}=-s'EI/L^3 \tag{15a}$$
$$M_{11}=-M_{23}=\bar{s}EI/L^2 \quad , \quad M_{21}=-M_{13}=\bar{s}EI/L^2 \tag{15b}$$
$$M_{a1}=-M_{b3}=\bar{s}EI/L^2 \quad , \quad M_{b1}=-M_{a3}=\bar{s}EI/L^2 \tag{15c}$$

2.5 Case 5

For Case 5 as shown in Fig.1, by imposing the boundary conditions $\theta''_a=1$, $\theta''_b=0$, and $v(0)=v'(0)=v(L)=v'(L)=0$ on Eq. (2), solution can be obtained via direct comparison of Case 2 and Case 5, where a unit plastic rotation at hinge '*a*' gives the same displacement pattern as a unit rotation at the '1' end. It follows that the forces and moments at the four DOFs and two PHLs are the same for both cases, i.e.,

$$V_{1a}=V_{12}=\bar{s}EI/L^2 \quad , \quad V_{2a}=V_{22}=-\bar{s}EI/L^2 \tag{16a}$$
$$M_{1a}=M_{12}=\hat{s}EI/L \quad , \quad M_{2a}=M_{22}=\hat{s}\hat{c}EI/L \tag{16b}$$
$$M_{aa}=M_{a2}=\hat{s}EI/L \quad , \quad M_{ba}=M_{b2}=\hat{s}\hat{c}EI/L \tag{16c}$$

2.6 Case 6

Finally, for Case 6 as shown in Fig.1, by imposing the boundary conditions $\theta''_a=0$, $\theta''_b=1$, and $v(0)=v'(0)=v(L)=v'(L)=0$ on Eq. (2), solution can be obtained via direct comparison of Case 4 and Case 6, where a unit plastic rotation at hinge '*b*' gives the same displacement pattern as a unit rotation at the '2' end. It follows that the forces and moments at the four DOFs and two PHLs are the same for both cases, i.e.,

$$V_{1b}=V_{14}=\bar{s}EI/L^2 \quad , \quad V_{2b}=V_{24}=-\bar{s}EI/L^2 \tag{17a}$$
$$M_{1b}=M_{14}=\hat{s}\hat{c}EI/L \quad , \quad M_{2b}=M_{24}=\hat{s}EI/L \tag{17b}$$
$$M_{ab}=M_{a4}=\hat{s}\hat{c}EI/L \quad , \quad M_{bb}=M_{b4}=\hat{s}EI/L \tag{17c}$$

2.7 Stiffness Matrices

In summary, based on Eqs. (6), (8), (11), (13), and (14)-(17) for the above six cases, the small-displacement-based stiffness matrix of the i^{th} member $\mathbf{k}_i^{SF}$ for bending while considering both geometric and material nonlinearities becomes

$$\mathbf{k}_i^{SF}=\frac{EI}{L^3}\begin{array}{c}\left[\begin{array}{cccc:cc} s' & \bar{s}L & -s' & \bar{s}L & \bar{s}L & \bar{s}L \\ \bar{s}L & \hat{s}L^2 & -\bar{s}L & \hat{s}\hat{c}L^2 & \hat{s}L^2 & \hat{s}\hat{c}L^2 \\ -s' & -\bar{s}L & s' & -\bar{s}L & -\bar{s}L & -\bar{s}L \\ \bar{s}L & \hat{s}\hat{c}L^2 & -\bar{s}L & \hat{s}L^2 & \hat{s}\hat{c}L^2 & \hat{s}L^2 \\ \hdashline \bar{s}L & \hat{s}L^2 & -\bar{s}L & \hat{s}\hat{c}L^2 & \hat{s}L^2 & \hat{s}\hat{c}L^2 \\ \bar{s}L & \hat{s}\hat{c}L^2 & -\bar{s}L & \hat{s}L^2 & \hat{s}\hat{c}L^2 & \hat{s}L^2 \end{array}\right]\end{array}\begin{array}{l} \leftarrow v(0) \\ \leftarrow v'(0) \\ \leftarrow v(L) \\ \leftarrow v'(L) \\ \leftarrow \theta''_a \\ \leftarrow \theta''_b \end{array} \tag{18}$$

where the superscript '*SF*' denotes the member stiffness matrix $\mathbf{k}_i$ is formulated by using the stability functions method that is computed based on the stability coefficients in Eqs. (7) and (12). Note again that $\mathbf{k}_i$ is 6×6, where the 6 movements are associated with the displacements and rotations at the two ends (i.e., 4 DOFs) and the plastic rotations at the two plastic hinge locations (i.e., 2 PHLs).

Linearization of Eq. (18) can be performed by using Taylor series expansion on each term of the member stiffness matrix and truncating higher-order terms. Doing so gives

$$\mathbf{k}_i^{GS}=\frac{EI}{L^3}\begin{bmatrix}12 & 6L & -12 & 6L\\ 6L & 4L^2 & -6L & 2L^2\\ -12 & -6L & 12 & -6L\\ 6L & 2L^2 & -6L & 4L^2\end{bmatrix}-\begin{array}{cl}\begin{bmatrix}6P/5L & P/10 & -6P/5L & P/10\\ P/10 & 2PL/15 & -P/10 & -PL/30\\ -6P/5L & -P/10 & 6P/5L & -P/10\\ P/10 & -PL/30 & -P/10 & 2PL/15\end{bmatrix} & \begin{array}{l}\leftarrow v(0)\\ \leftarrow v'(0)\\ \leftarrow v(L)\\ \leftarrow v'(L)\end{array}\end{array}\qquad(19)$$

where the first matrix in Eq. (19) represents that classic stiffness matrix without considering any geometric nonlinearity, and the second matrix represents the geometric stiffness. The superscript ‘*GS*’ denotes the member stiffness matrix $\mathbf{k}_i$ is formulated by using the geometric stiffness method. Note that the member stiffness matrix in Eq. (19) is a 4×4 matrix, where the rows and columns associated with θ''_a and θ''_b are dropped. This is because solution algorithms among various software packages that incorporate geometric nonlinearity using the geometric stiffness method usually adopt a different and independent algorithm for material nonlinearity.

Finally, the member stiffness matrix $\mathbf{k}_i$ in Eq. (19) can be further simplified by retaining only the large *P*-Δ effect while ignoring small *P*-δ effect. This is done by removing all geometric nonlinear terms associated with bending. Doing so gives

$$\mathbf{k}_i^{PD}=\frac{EI}{L^3}\begin{bmatrix}12 & 6L & -12 & 6L\\ 6L & 4L^2 & -6L & 2L^2\\ -12 & -6L & 12 & -6L\\ 6L & 2L^2 & -6L & 4L^2\end{bmatrix}-\begin{array}{cl}\begin{bmatrix}P/L & 0 & -P/L & 0\\ 0 & 0 & 0 & 0\\ -P/L & 0 & P/L & 0\\ 0 & 0 & 0 & 0\end{bmatrix} & \begin{array}{l}\leftarrow v(0)\\ \leftarrow v'(0)\\ \leftarrow v(L)\\ \leftarrow v'(L)\end{array}\end{array}\qquad(20)$$

where the superscript ‘*PD*’ denotes the member stiffness matrix $\mathbf{k}_i$ is formulated by using the *P*-Δ stiffness method.

3. Identification of Geometric Nonlinearity Using a Single Degree of Freedom System

An appropriate model should be used to serve the purpose of the study. To identify the geometric nonlinearity used in different software packages, consider the single degree of freedom (SDOF) column model as shown in Fig.2a be subjected to a constant axial compressive force *P*. Let the mass be *M*, damping be *C*, elastic modulus be *E*, moment of inertia be *I*, and length be *L*. The degree of freedom is set up so that the mass is restrained from rotation, but it is free to translate in the horizontal direction.

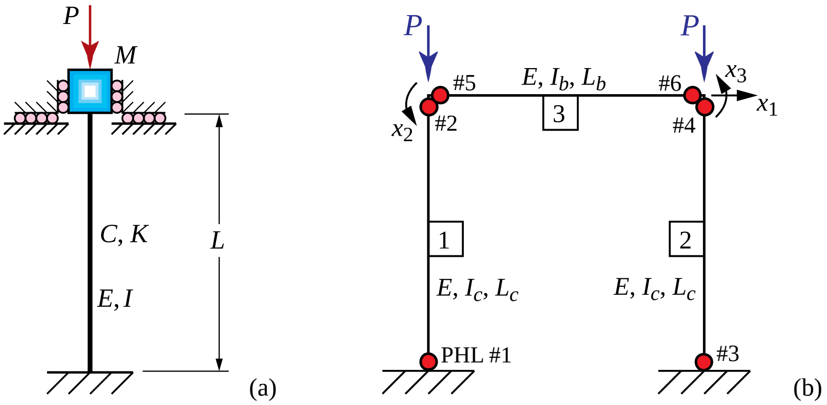


Fig. 2 – Single degree of freedom models with geometric nonlinearity

Based on this set up, the column experiences a constant axial force due to the gravity load only and also experiences shear and moment due to the earthquake ground motion only. Therefore, the lateral stiffness of the column remains constant with a constant axial force throughout the entire earthquake time history analysis. If geometric nonlinearity is totally ignored in this problem, the lateral stiffness related to the horizontal translation degree of freedom at the top of the column is simply:

$$K_{OR} = \frac{12EI}{L^3} \tag{21}$$

where K_{OR} represents the original stiffness of the SDOF system without considering any geometric nonlinearity. If geometric nonlinearity is considered using the *P*-Δ stiffness approach, the lateral stiffness of the SDOF column (see Eq. (20)) becomes:

$$K_{P\Delta} = \frac{12EI}{L^3} - \frac{P}{L} \tag{22}$$

where $K_{P\Delta}$ denotes the geometrically nonlinear stiffness of the column using the *P*-Δ approach (i.e., including large *P*-Δ effect but excluding small *P*-δ effect). If geometric nonlinearity is considered using the geometric stiffness approach, the lateral stiffness of the SDOF column (see Eq. (19)) becomes:

$$K_{GS} = \frac{12EI}{L^3} - \frac{6P}{5L} \tag{23}$$

where K_{GS} denotes the geometrically nonlinear stiffness of the column using the geometric stiffness approach (i.e., including both large *P*-Δ and small *P*-δ effects). Finally, if geometric nonlinearity is considered through the use of stability functions, the stiffness matrix (see Eq. (18)) becomes:

$$K_{SF} = \frac{s'EI}{L^3} \tag{24}$$

where K_{SF} denotes the geometrically nonlinear stiffness computed using the stability functions approach (i.e., including both large *P*-Δ and small *P*-δ effects in a consistent form), and s' is the stability coefficients defined in Eq. (12).

In this study, let $E = 100$ GPa, $I = 6.4\times10^5$ mm^4, $L = 4$ m, and $P = 4$ kN in compression. Using a mass of $M = 9500$ kg, the calculated stiffnesses and the corresponding periods of vibration are summarized in Table 1. Note that the critical buckling load for the SDOF column is $P_{cr} = \pi^2 EI/L^2 = 39.48$ kN, which means the applied load is at 10.1 % of the critical buckling load (i.e., $P / P_{cr} = 0.101$). At this axial compressive force level, it is observed that $K_{GS} \approx K_{SF}$, but there is a 1.9 % difference in stiffness between $K_{P\Delta}$ and K_{GS}. This suggests that ignoring small *P*-δ effect can result in an increase in lateral stiffness by 1.9 %. Table 1 also shows that the difference between using $K_{P\Delta}$ and using K_{GS} results in a 1.0 % difference in the calculated period of vibration of the SDOF column. Note that 1.0 % elongation of period is sufficient to provide the phase shift needed to identify the geometric nonlinearity used in software packages.

Table 1 – Comparison of geometric nonlinear stiffnesses, periods, and maximum responses

	Stiffness (kN/m)	Period (s)	Max Displacement (m)	Max Velocity (m/s)	Max Acceleration (*g*)
K_{OR}	12.0	1.496	0.6990	2.996	1.258
$K_{P\Delta}$	11.0	1.562	0.6737	2.789	1.111
K_{GS}	10.8	1.577	0.6625	2.746	1.073
K_{SF}	10.799	1.577	0.6624	2.746	1.073

By assuming 0 % damping, the SDOF column shown in Fig.2a is now subjected to the 1995 Kobe earthquake as shown in Fig.3. The resulting displacement responses using the original stiffness (OR), *P*-Δ stiffness (PD), geometric stiffness (GS), and stability function stiffness (SF) are presented in Fig.4. In addition, the maximum displacement, velocity, and absolute acceleration for using various geometric nonlinear stiffness approaches are summarized in Table 1. It can be seen from Fig. 4 that the responses using K_{GS} match those using K_{SF} well at such a small axial compressive force. The differences are more noticeable between the responses using $K_{P\Delta}$ and those using K_{GS}, where the maximum displacement response using $K_{P\Delta}$ increases by 1.7 % and the maximum acceleration response also increases by 3.5 %.

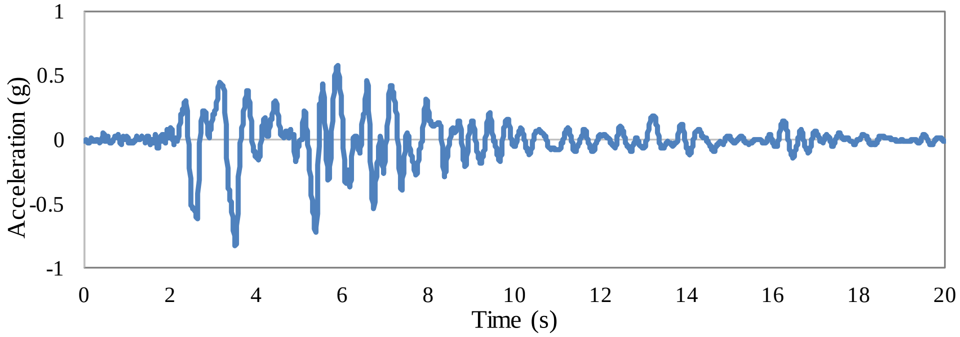


Fig. 3 – 1995 Kobe earthquake ground acceleration time history

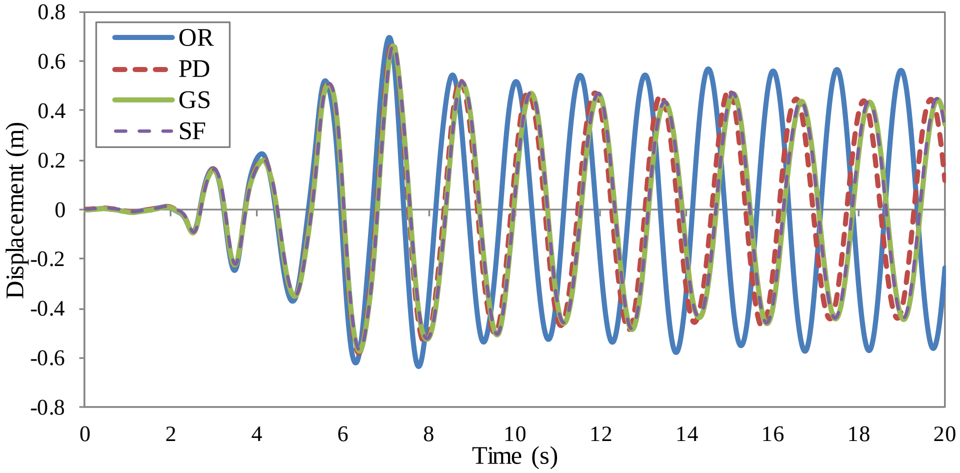


Fig. 4 – Displacement responses of the SDOF column using different stiffnesses

Knowing the stiffness of the column is important because it can be used to assess the type of geometric nonlinearity that is embedded in various software packages. Consider the same SDOF column shown in Fig.2a is now modeled using four small-displacement-based software packages commonly used in the United States, randomly labeled as S1, S2, S3, and S4, and is subjected to the 1995 Kobe earthquake shown in Fig.3. By assuming 0 % damping, Fig.5 shows the displacement responses from these software packages, and these responses are compared with those in Fig.4. As shown in Fig.5a, the comparison shows an exact match between the current analysis method using K_{GS} and the software packages S1 and S4. Similarly, Fig.5b shows an exact match between the current analysis method with $K_{P\Delta}$ and software packages S2 and S3. This indicates two of the small-displacement software packages use geometric stiffness that considers both large P-Δ and small P-δ effects in the formulation, while the other two of the small-displacement software packages use P-Δ stiffness that considers only large P-Δ effects and ignores the small P-δ effects. Note that even though using 0 % damping is an idealized situation, it helps eliminate the potentially differing effects of using damping parameters on the responses that may occur due to differences in damping formulations used in different software packages.

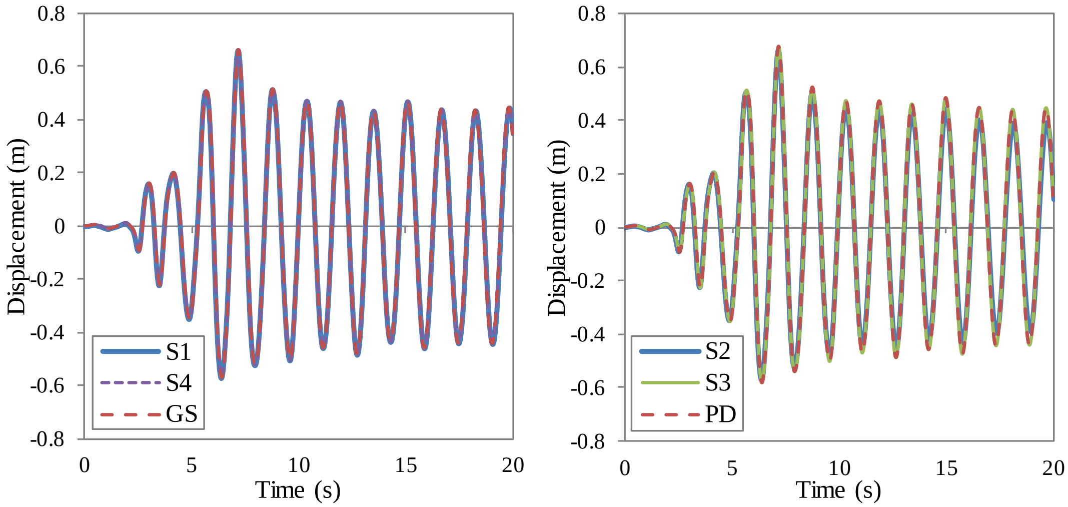


Fig. 5 – Displacement response comparisons between various software packages and manual calculations

4. One-Story Moment Frame

Once the type of geometric nonlinearity is known, the influence of using different types of geometric nonlinearity on displacement responses can be assessed. As a simple numerical example, consider a one-story one-bay moment-resisting frame as shown in Fig.2b with members assumed to be axially rigid. This gives a total of 3 DOFs (i.e., $n = 3$) and 6 PHLs (i.e., $m = 6$) as shown in the figure. Software packages S1 (based on GS) and S2 (based on PD) are selected for this study along with the SF method. Since using SF formulation is an uncommon approach with limited documentations, it is therefore worthwhile to discuss the analysis approach here. Based on Eq. (18), the member stiffness matrix for the two columns can be written as:

$$\mathbf{k}_i = \frac{EI_c}{L_c^3}\left[\begin{array}{cccc|cc} s'_i & \bar{s}_i L_c & -s'_i & \bar{s}_i L_c & \bar{s}_i L_c & \bar{s}_i L_c \\ \bar{s}_i L_c & \hat{s}_i L_c^2 & -\bar{s}_i L_c & \hat{s}_i \hat{c}_i L_c^2 & \hat{s}_i L_c^2 & \hat{s}_i \hat{c}_i L_c^2 \\ -s'_i & -\bar{s}_i L_c & s'_i & -\bar{s}_i L_c & -\bar{s}_i L_c & -\bar{s}_i L_c \\ \bar{s}_i L_c & \hat{s}_i \hat{c}_i L_c^2 & -\bar{s}_i L_c & \hat{s}_i L_c^2 & \hat{s}_i \hat{c}_i L_c^2 & \hat{s}_i L_c^2 \\ \hline \bar{s}_i L_c & \hat{s}_i L_c^2 & -\bar{s}_i L_c & \hat{s}_i \hat{c}_i L_c^2 & \hat{s}_i L_c^2 & \hat{s}_i \hat{c}_i L_c^2 \\ \bar{s}_i L_c & \hat{s}_i \hat{c}_i L_c^2 & -\bar{s}_i L_c & \hat{s}_i L_c^2 & \hat{s}_i \hat{c}_i L_c^2 & \hat{s}_i L_c^2 \end{array}\right] , \quad i = 1, 2 \tag{25}$$

where $\hat{s}_i$, $\hat{c}_i$, $\bar{s}_i$, and s'_i are stability coefficients of the i^{th} column member computed using Eqs. (7) and (12). No axial force is acting on the beam member, and therefore the beam member stiffness matrix becomes:

$$\mathbf{k}_3 = \frac{EI_b}{L_b^3}\left[\begin{array}{cccc|cc} 12 & 6L_b & -12 & 6L_b & 6L_b & 6L_b \\ 6L_b & 4L_b^2 & -6L_b & 2L_b^2 & 4L_b^2 & 2L_b^2 \\ -12 & -6L_b & 12 & -6L_b & -6L_b & -6L_b \\ 6L_b & 2L_b^2 & -6L_b & 4L_b^2 & 2L_b^2 & 4L_b^2 \\ \hline 6L_b & 4L_b^2 & -6L_b & 2L_b^2 & 4L_b^2 & 2L_b^2 \\ 6L_b & 2L_b^2 & -6L_b & 4L_b^2 & 2L_b^2 & 4L_b^2 \end{array}\right] \tag{26}$$

Assembling the above member stiffness matrices into the global stiffness matrix gives a 9×9 stiffness matrix, which can be partitioned into the form

$$\mathbf{K}^{SF}=\left[\begin{array}{c|c}\mathbf{K}_{n\times n} & \mathbf{K}'_{n\times m}\\ \hline \mathbf{K}'^{T}_{m\times n} & \mathbf{K}''_{m\times m}\end{array}\right]=\left[\begin{array}{c|c}\mathbf{K}_{3\times 3} & \mathbf{K}'_{3\times 6}\\ \hline \mathbf{K}'^{T}_{6\times 3} & \mathbf{K}''_{6\times 6}\end{array}\right]\begin{array}{l}\leftarrow x_1,x_2,x_3\\ \leftarrow \theta''_1,\ldots,\theta''_6\end{array} \tag{27}$$

where

$$\mathbf{K}_{3\times 3}=\begin{bmatrix}(s'_1+s'_2)EI_c/L_c^3 & \bar{s}_1EI_c/L_c^2 & \bar{s}_2EI_c/L_c^2\\ \bar{s}_1EI_c/L_c^2 & \hat{s}_1EI_c/L_c+4EI_b/L_b & 2EI_b/L_b\\ \bar{s}_2EI_c/L_c^2 & 2EI_b/L_b & \hat{s}_2EI_c/L_c+4EI_b/L_b\end{bmatrix}\begin{array}{l}\leftarrow x_1\\ \leftarrow x_2\\ \leftarrow x_3\end{array} \tag{28a}$$

$$\mathbf{K}'_{3\times 6}=\left[\begin{array}{cc|cc|cc}\bar{s}_1EI_c/L_c^2 & \bar{s}_1EI/L_c^2 & \bar{s}_2EI_c/L_c^2 & \bar{s}_2EI_c/L_c^2 & 0 & 0\\ \hat{s}_1\hat{c}_1EI_c/L_c & \hat{s}_1EI_c/L_c & 0 & 0 & 4EI_b/L_b & 2EI_b/L_b\\ 0 & 0 & \hat{s}_2\hat{c}_2EI_c/L_c & \hat{s}_2EI_c/L_c & 2EI_b/L_b & 4EI_b/L_b\end{array}\right]\begin{array}{l}\leftarrow x_1\\ \leftarrow x_2\\ \leftarrow x_3\end{array} \tag{28b}$$

$$\mathbf{K}''_{6\times 6}=\left[\begin{array}{cc|cc|cc}\hat{s}_1EI_c/L_c & \hat{s}_1\hat{c}_1EI_c/L_c & 0 & 0 & 0 & 0\\ \hat{s}_1\hat{c}_1EI_c/L_c & \hat{s}_1EI_c/L_c & 0 & 0 & 0 & 0\\ \hline 0 & 0 & \hat{s}_2EI_c/L_c & \hat{s}_2\hat{c}_2EI_c/L_c & 0 & 0\\ 0 & 0 & \hat{s}_2\hat{c}_2EI_c/L_c & \hat{s}_2EI_c/L_c & 0 & 0\\ \hline 0 & 0 & 0 & 0 & 4EI_b/L_b & 2EI_b/L_b\\ 0 & 0 & 0 & 0 & 2EI_b/L_b & 4EI_b/L_b\end{array}\right]\begin{array}{l}\leftarrow \theta''_1\\ \leftarrow \theta''_2\\ \leftarrow \theta''_3\\ \leftarrow \theta''_4\\ \leftarrow \theta''_5\\ \leftarrow \theta''_6\end{array} \tag{28c}$$

Assume that the frame shown in Fig.2b has a mass of 318.7 Mg and a damping of 0 %. Also, let $E=200$ GPa, $I_b=I_c=4.995\times10^8$ mm^4, $L_b=7.62$ m, $L_c=4.57$ m, and $P=5{,}338$ kN. To include material nonlinearity in the analysis, assume that all six plastic hinges exhibit elastic-plastic behavior with plastic moment capacities of $m_b=3130$ kN·m for the beam and $m_c=3909$ kN·m for the two columns. By subjecting the frame to the 1995 Kobe earthquake ground motion as shown in Fig.3 with a scale factor of 1.3, the global displacement response at DOF #1 (i.e., roof displacement) is plotted in Fig.6 based on the stiffness matrices using stability function formulations (SF) in Eq. (28). In addition, the same undamped responses obtained from the software package S1 that uses geometric stiffness (GS) and the software package S2 that uses P-Δ stiffness (PD) are plotted in the figure for comparison. It is observed that the choice of geometric nonlinearity used in the analysis can influence the response calculations even for a simple SDOF system. While the responses calculated using SF and GS are reasonably close to each other, the calculated response using PD is quite different from the other two calculated responses.

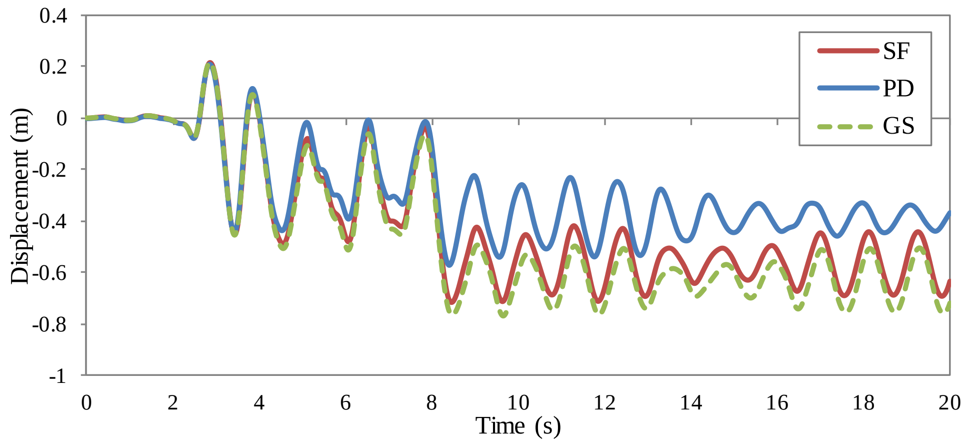


Fig. 6 – Displacement response comparisons of a one-story frame with various geometric nonlinearities

5. Four-Story Moment-Resisting Frame

Consider the four-story moment-resisting steel frame as shown in Fig.7a. This frame contains 36 DOFs (i.e., $n = 36$) and 56 PHLs (i.e., $m = 56$). Assume a mass of 72 670 kg is used on each floor, and no mass or mass moment of inertia is assigned to any of the vertical translation DOFs nor rotation DOFs. A gravity load of 863 kN is applied on each exterior column member and 1263 kN is applied on each interior column member as shown in Fig.7b. In addition, 0 % damping is assumed in order to provide a better comparison of the response among different software packages, including software package S2 that uses *P*-Δ stiffness (PD) and LS-DYNA finite element software package (LD) that uses large displacement formulation. For stability function (SF), the global stiffness matrix in Eq. (27) after assembly includes $\mathbf{K}_{36\times36}$, $\mathbf{K}'_{36\times56}$, and $\mathbf{K}''_{56\times56}$.

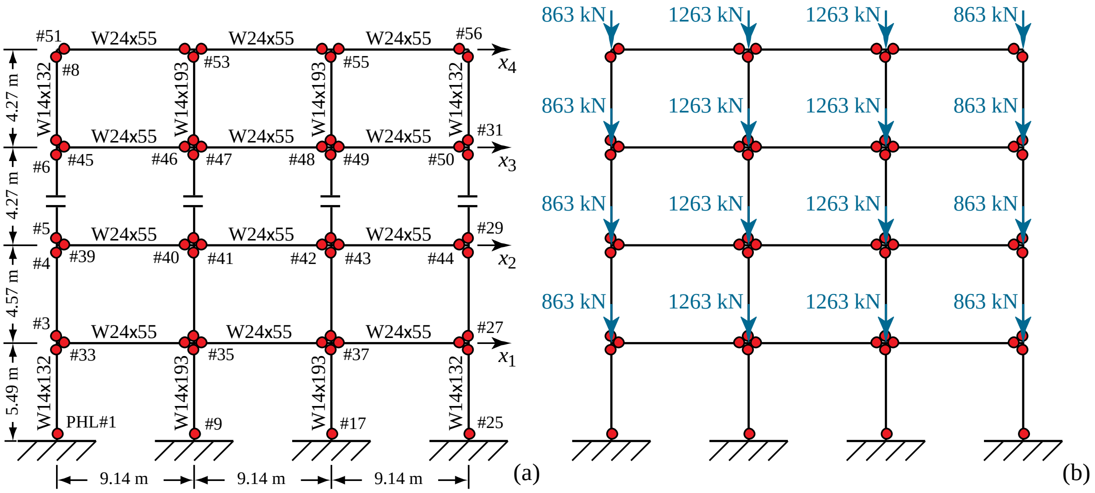


Fig. 7 – Four-story moment-resisting steel frame with gravity loads

Assume the yield stress of the member is 345 MPa and all 56 plastic hinges exhibit elastic-plastic behavior. The steel frame is now subjected to the 1995 Kobe earthquake ground motion as shown in Fig.3 with scale factors of 0.6, 0.8, 1.0, and 1.2. The roof displacement responses are summarized in Fig.8. Based on the results, it is observed that the responses computed using different geometric nonlinearity are similar at a scale factor of 0.6 when the response is only slightly nonlinear. However, the PD response begins to deviate at a scale factor of 0.8 with significant differences at a scale factor of 1.0. This suggests that using PD software packages that captures only large *P*-Δ effect while ignoring small *P*-δ effects may not be appropriate for computation when the significant coupling between geometric nonlinearity and material nonlinearity is expected.

6. Conclusion

In this paper, different formulations of geometric nonlinearity were presented with a detailed derivation of the stability functions stiffness matrix while incorporating material nonlinearity. Through this derivation, how each formulation captures large *P*-Δ and small *P*-δ effects were explained, and the investigation continued with a study on how different software packages implement geometric nonlinearity. A SDOF column was modeled using four different small-displacement-based software packages. Based on the output dynamic responses it was identified that two packages use *P*-Δ stiffness that includes only large *P*-Δ effects while the other two packages use geometric stiffness that includes both large *P*-Δ and small *P*-δ effects. The study was then extended to see how different formulations of geometric nonlinearity impact the response of moment-resisting frame, and it was shown that software packages implementing *P*-Δ stiffness can lead to different calculated results when there is significant coupling between geometric and material nonlinearities.

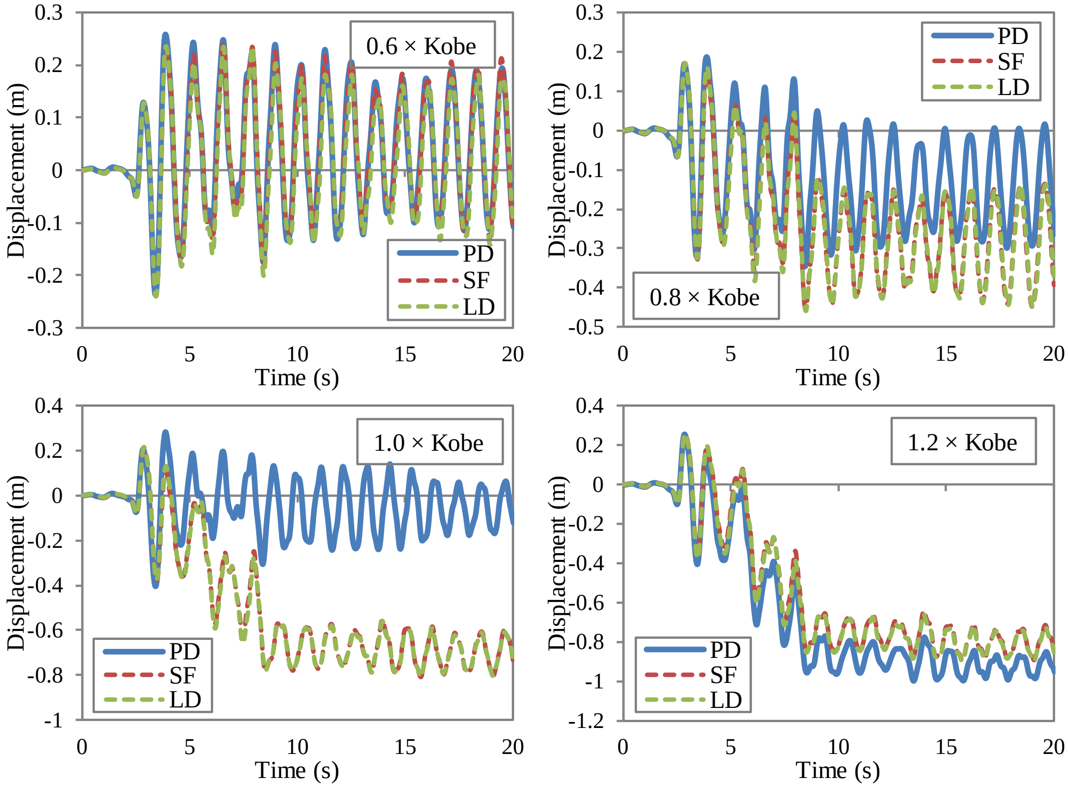


Fig. 8 – Roof displacement response of the four-story frame with various geometric nonlinearities

Since both ASCE/SEI 7-16 and ASCE/SEI 41-17 are design standards with the purpose of limiting damages in structures during major earthquakes, only slight material nonlinearity is expected in the analysis of the designed models. Therefore, any formulation of geometric nonlinearity is appropriate for use in these standards as long as users 'turn on' geometric nonlinearity in the analysis. However, if the analysis requires significant coupling between geometric and material nonlinearities, such as in the case of analyzing structural collapse or near-collapse, the applicability of software packages that implement *P*-Δ stiffness warrants further study.

7. References

[1] ASCE/SEI 7-16 (2016): *Minimum Design Loads for Buildings and Other Structures*. American Society of Civil Engineers, Reston, VA, USA.

[2] ASCE/SEI 41-17 (2017): *Seismic Evaluation and Retrofit of Existing Buildings*. American Society of Civil Engineers, Reston, VA, USA.

[3] Timoshenko SP, Gere JM (1961): *Theory of Elastic Stability*. McGraw Hill, 2nd Edition, NY, USA.

[4] Bazant ZP, Cedolin L (2003): *Stability of Structures*. Dover Publication, NY, USA.

[5] Powell GH (2010): *Modeling for Structural Analysis: Behavior and Basics*. Computers and Structures, CA, USA.

[6] Wilson E (2010): *Static and Dynamic Analysis of Structures: A Physical Approach with Emphasis on Earthquake Engineering*. Computer and Structures, 4th Edition, CA, USA.

Combinatorial Rank Attacks Against the Rectangular Simple Matrix Encryption Scheme

Daniel Apon[1], Dustin Moody[1], Ray Perlner[1], Daniel Smith-Tone[1,2], and Javier Verbel[3]

[1]National Institute of Standards and Technology, USA
[2]University of Louisville, USA
[3]Universidad Nacional de Colombia, Colombia

daniel.apon@nist.gov,dustin.moody@nist.gov, ray.perlner@nist.gov, daniel.smith@nist.gov, javerbel@unal.edu.co

Abstract. In 2013, Tao et al. introduced the ABC Simple Matrix Scheme for Encryption, a multivariate public key encryption scheme. The scheme boasts great efficiency in encryption and decryption, though it suffers from very large public keys. It was quickly noted that the original proposal, utilizing square matrices, suffered from a very bad decryption failure rate. As a consequence, the designers later published updated parameters, replacing the square matrices with rectangular matrices and altering other parameters to avoid the cryptanalysis of the original scheme presented in 2014 by Moody et al.

In this work we show that making the matrices rectangular, while decreasing the decryption failure rate, actually, and ironically, diminishes security. We show that the combinatorial rank methods employed in the original attack of Moody et al. can be enhanced by the same added degrees of freedom that reduce the decryption failure rate. Moreover, and quite interestingly, if the decryption failure rate is still reasonably high, as exhibited by the proposed parameters, we are able to mount a reaction attack to further enhance the combinatorial rank methods. To our knowledge this is the first instance of a reaction attack creating a significant advantage in this context.

Keywords: Multivariate Cryptography, Simple Matrix, encryption, MinRank

1 Introduction

Since the discovery by Peter Shor in the 1990s, cf. [26], of polynomial-time quantum algorithms for computing discrete logarithms and factoring integers the proverbial clock has been ticking on our current public key infrastructure. In reaction to this discovery and the continual advancement of quantum computing technologies, a large community has emerged dedicated to the development and deployment of cryptosystems that are immune to the exponential speedups quantum computers promise for our current standards. More recently, the National

Moody, Dustin; Perlner, Ray; Smith-Tone, Daniel; Apon, Daniel; Verbel, Javier. "Combinatorial Rank Attacks Against the Rectangular Simple Matrix Encryption Scheme." Paper presented at PQCrypto 2020: The Eleventh International Conference on Post-Quantum Cryptography, Paris, FR. April 15, 2020 - April 17, 2020.

Institute of Standards and Technology (NIST) has begun directing a process to reveal which of the many new options for post-quantum public key cryptography are suitable for widespread use.

One family of candidate schemes relies on the known difficulty of solving large systems of nonlinear equations. These multivariate public key cryptosystems are inspired by computational problems that have been studied by algebraic geometers for several decades. Still, even in the past two decades this field of study has changed dramatically.

When multivariate public key cryptography was still early in its community building phase, a great many schemes were proposed and subsequently attacked. Notable examples of this phenomenon include C^*, HFE, STS, Oil-Vinegar, PMI and SFLASH, see [14, 22, 32, 19, 6, 20, 21, 9, 25, 10, 8].

While multivariate cryptography has seen some lasting success with digital signatures, see, for example, [12, 4, 2, 23, 5], multivariate encryption seems to be particularly challenging. In the last several years there have been many new proposals inspired by the notion that it may be easier to create a secure injective multivariate function if the codomain is larger than the domain. Such schemes include ZHFE, Extension Field Cancellation (EFC), SRP, HFERP, EFLASH and the Simple Matrix Encryption Scheme, see [7, 28, 34, 11, 3, 29, 30]. Of these, many have since endured attacks either outright breaking the scheme or affecting parameters, see [1, 27, 24, 15–17].

In this work we present a new attack on the rectangular variant of the Simple Matrix Encryption Scheme, see [30]. This version of the Simple Matrix Encryption Scheme was designed to repair the problems that the original scheme, see [29], had with decryption failures and to choose large enough fields to avoid the attack of [15]. Our new attack is still a MinRank method, but one that exploits the rectangular structure, showing that the new parameterization is actually less secure than the square variant.

In an interesting twist, we also develop a reaction attack based on the decryption failures that the scheme is designed to minimize. This method further boosts the performance of the MinRank step by a factor related to the field size. With these attacks we break all of the published parameter sets at the most efficient field size of 2^8, the only parameters for which performance data were offered.

The article is organized as follows. In Section 2, we present the Simple Matrix Scheme. We next review the MinRank attack techniques using properties of the differential that was used against the original square variant of the Simple Matrix scheme. In the subsequent section, we present the improvement obtained in attacking the rectangular variant. Next, in Section 5, we present the reaction attack and discuss its affect on key recovery. We then present a thorough complexity analysis including our experimental data verifying our claimed complexity. Finally, we conclude noting the effect this attack has on the status of multivariate encryption.

2 ABC Simple Matrix Scheme

The ABC Simple Matrix Encryption Scheme was introduced in [29] by Tao et al. This scheme was designed with a new guiding principle in mind: make the codomain much larger than the domain. The motivation for this notion comes from the fact that there is a much richer space of injective functions with a large codomain than the space of bijective functions; thus, it may be easier to hide the types of properties we use to efficiently invert nonlinear functions such as low rank or low degree in this larger context. In this section we present the scheme and its functionality.

For clarity of exposition, we establish our notational standard. Throughout this text bold font will indicate a matrix or vector, e.g. **T** or **z**, while regular fonts indicate functions (possibly with outputs considered as matrices) or field elements.

2.1 ABC Public Key Generation

Let $\mathbb{F}$ be a finite field with q elements. Let s be a positive integer and let $n = s^2$. Let $\mathbb{F}[\mathbf{x}]$ be the polynomial ring over $\mathbb{F}$ in the variables $\mathbf{x} = \begin{bmatrix} x_1 \cdots x_n \end{bmatrix}$.

The public key will be a system of $m = 2n = 2s^2$ (for our purposes homogeneous) quadratic formulae in $\mathbb{F}[\mathbf{x}]$. The public key will ultimately be generated by the standard isomorphism construction $P = T \circ F \circ U$ where T and U are invertible linear transformations of the appropriate dimensions, and F is a specially structured system of quadratic polynomials. The remainder of this section is devoted to the construction of F. (In general the scheme can and does use rectangular matrices, but for the ease of writing this note, we will assume that the matrices are square for now.)

Define the matrix

$$\mathbf{A} = \begin{bmatrix} x_1 & \cdots & x_s \\ x_{s+1} & \cdots & x_{2s} \\ \vdots & \ddots & \vdots \\ x_{s^2-s+1} & \cdots & x_{s^2} \end{bmatrix}.$$

Further define the $s \times s$ matrices of $\mathbb{F}[\mathbf{x}]$ linear forms $\mathbf{B} = \begin{bmatrix} b_{ij} \end{bmatrix}$ and $\mathbf{C} = \begin{bmatrix} c_{ij} \end{bmatrix}$.

From these matrices one can construct the matrices $\mathbf{E}_1 = \mathbf{AB}$ and $\mathbf{E}_2 = \mathbf{AC}$. Then we construct a system of m polynomials by concatenating the vectorizations of these two products: $F = Vec(\mathbf{E}_1) \| Vec(\mathbf{E}_2)$. The public key is then $P = T \circ F \circ U$. (Note that we can eliminate U by replacing $\mathbf{A}$ with random linear forms.)

In the rectangular version of this scheme we replace $\mathbf{A}$ by a similar $r \times s$ version (and we can make the matrices $\mathbf{B}$ and $\mathbf{C}$ of size $s \times u$ and $s \times v$, respectively) where the algebra still works the same.

2.2 Encryption and Decryption

Encryption is accomplished by evaluating the public key at a plaintext value encoded as a vector $\mathbf{x}$. One computes $P(\mathbf{x}) = \mathbf{y}$.

Decryption is accomplished by inverting each of the components of the public key. One first sets $\mathbf{v} = T^{-1}(\mathbf{y})$. Then $\mathbf{v}$ can be split in half producing $\mathbf{v}_1$ and $\mathbf{v}_2$. Each of these can be parsed as a matrix by inverting the vectorization operator $\mathbf{E}_1 = Mat(\mathbf{v}_1)$ and $\mathbf{E}_2 = Mat(\mathbf{v}_2)$.

We note that we can consider this pair of matrices as values derived from functions on either the inputs $\mathbf{x}$ or the outputs $\mathbf{y}$. The legitimate user knows both of these representations. We will abuse notation slightly and denote these functions as $E_1(\mathbf{u})$, $E_1(\mathbf{v})$, $E_2(\mathbf{u})$ and $E_2(\mathbf{v})$, where $\mathbf{v} = F(\mathbf{u})$ (and we use similar notation for functions of $\mathbf{u}$ representing the matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$. Thus, we have computed $\mathbf{E}_1 = E_1(\mathbf{v})$ and $\mathbf{E}_2 = E_2(\mathbf{v})$. These values must be equal to $E_i(\mathbf{u})$. For both values of i, the function involves a left product by the square matrix $A(\mathbf{u})$. We construct a matrix $\mathbf{W}$ of new variables w_i for $0 < i \leq s^2$. We suppose that the correct assignment of values in $A(\mathbf{u})$ produces a matrix with a left inverse, so the correct assignment of variables w_i produces a valid left inverse. Then we have

$$\mathbf{W}\mathbf{E}_1 = \mathbf{W}E_1(\mathbf{u}) = \mathbf{W}A(\mathbf{u})B(\mathbf{u}) = B(\mathbf{u}),$$

and similarly for $\mathbf{E}_2$. Since the legitimate user knows the linear forms b_{ij} and c_{ij}, this setup provides a system of $m = 2s^2$ equations in the $s^2 + s^2$ variables w_i and u_i. Via Gaussian elimination, the w_i variables can be eliminated and values for u_i can be recovered.

Once $\mathbf{u}$ is recovered, one applies the inverse of U to this quantity to recover $\mathbf{x}$, the plaintext.

3 Previous Cryptanalysis

In this section, we summarize the technique from [15] recovering a secret key in the square case, that is when $r = s$, via MinRank informed by differential invariant structure. For convenience, we present the relevant definitions we will use in Section 4, possibly generalized to the rectangular setting.

The main object used in the attack from [15] is the discrete differential of the public key.

Definition 1 *Let $F : \mathbb{F}^n \to \mathbb{F}^m$. The* discrete differential *of F is a bivariate analogue of the discrete derivative; it is given by the normalized difference*

$$DF(\mathbf{a}, \mathbf{x}) = F(\mathbf{a} + \mathbf{x}) - F(\mathbf{a}) - F(\mathbf{x}) + F(\mathbf{0}).$$

DF is a vector-valued function since the output is in $\mathbb{F}^m$. Since DF is bilinear, we can think of each coordinate DF_i as a matrix. We can then consider properties of these matrices as linear operators. In particular, we can consider rank and perform a MinRank attack.

Definition 2 *The* MinRank(q, n, m, r) *Problem is the task of finding a linear combination over $\mathbb{F}_q$ of m matrices, $\mathbf{DQ}_i$, of size $n \times n$ such that the resulting rank is at most r.*

Although there are many different techniques for solving MinRank, the most relevant technique here is known as *linear algebra search*. One attempts to guess $\ell = \lceil \frac{m}{n} \rceil$ vectors that lie in the kernel of the same map. Since matrices with low rank have more linearly independent vectors in their kernels, the distribution of maps whose kernels contain these vectors is skewed toward lower rank maps. Therefore, to solve MinRank, one guesses ℓ vectors $\mathbf{x}_i$, sets up the linear system

$$\sum_{i=1}^{m} \tau_i \mathbf{D}\mathbf{Q}_i \mathbf{x}_j = \mathbf{0},$$

for $j = 1, \ldots, \ell$, solves for τ_i and computes the rank of $\sum_{i=1}^{m} \tau_i \mathbf{D}\mathbf{Q}_i$. If the rank is at or below the target rank then the attack has succeeded. Otherwise another set of vectors is chosen and the process continues.

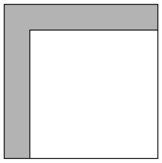
In [15], the attack is formulated in the language of differential invariants.

Definition 3 *A* subspace differential invariant *of a vector-valued map F is a triple of vector spaces (X, V, W) such that $X \subseteq \mathbb{F}^m$, and $V, W \subseteq \mathbb{F}^n$ satisfying $(\mathbf{x} \cdot DF)V \subseteq W$ for all $\mathbf{x} \in X$ where $dim(W) \leq dim(V)$.*

In other words, a subspace differential invariant is a subspace X of the span of the DF_i along with a subspace that is mapped linearly by every map in X into another subspace of no larger dimension. The definition is supposed to capture the idea of a subspace of the span of F acting like a linear map on a subspace of the domain of F.

Differential invariants are related to low rank, but not equivalent. They are useful at providing an algebraic condition on interlinked kernels, that is, when there are very many maps in the span of F that have low rank and share a large common subspace in their kernels, see [33]. In such a case, the invariant structure provides a tiny and insignificant savings in some linear algebra steps after the hard MinRank step of the attack is complete. The main value of the idea lies in providing algebraic tools for determining whether an interlinked kernel structure is present in a map.

Considering the Simple Matrix Scheme, there are maps in the span of the public maps that correspond to products of the first row of $\mathbf{A}$ and linear combinations of the columns of $\mathbf{B}$ and $\mathbf{C}$. The differential of this type of map has the following structure, where gray indicates possibly nonzero coefficients.

$Dg =$ 

This map is clearly of low rank, probably $2s$, and illustrates a differential invariant because a column vector with zeros in the top s entries is mapped by this matrix to a vector with zeros in everything except the top s entries. Also, it is important to note that there is an entire $u + v$ dimensional subspace of the public key corresponding to the X in Definition 3 that produces differentials of this shape which we call a band space. There is nothing special about the first row. We could use anything in the rowspace of $\mathbf{A}$ and express our differential as above in the appropriate basis. This motivates the following definition modified from [15, Definition 4]:

Definition 4 *Fix an arbitrary vector* $\mathbf{v}$ *in the rowspace of* $\mathbf{A}$*, i.e.* $\mathbf{v} = \sum_{d=1}^{r} \lambda_d \mathbf{A}_d$ *where* $\mathbf{A}_d$ *is the dth row of* $\mathbf{A}$*. The* $u + v$ *dimensional space of quadratic forms* $\mathcal{B}_v$ *given by the span of the columns of* $\mathbf{vB}$ *and* $\mathbf{vC}$ *is called the generalized band-space generated by* $\mathbf{v}$*.*

Thus, recovery of an equivalent private key is accomplished by discovering r linearly independent band spaces in the span of the public key. Since these maps all share the property that they are of rank $2s$, the band spaces can be recovered with a MinRank attack.

Due to the differential invariant structure, it is shown in [15] that there is a significant speed-up in the standard linear algebra search variant of MinRank. The attack proceeds by finding $\lceil \frac{m}{n} \rceil$ vectors in the kernel of the same band space map.

A series of statements about such maps are proven in [15] in the square case revealing the complexity of the MinRank step of the attack.

Definition 5 *Let* $u_1, \ldots, u_{rs}$ *be the components of* $\mathbf{Ux}$ *and fix an arbitrary vector* $\mathbf{v}$ *in the rowspace of* $\mathbf{A}$*, i.e.* $\mathbf{v} = \sum_{d=1}^{r} \lambda_d \mathbf{A}_d$ *where* $\mathbf{A}_d$ *is the dth row of* $\mathbf{A}$*. An* rs*-dimensional vector,* $\mathbf{x}$ *is in the band kernel generated by* $\mathbf{v}$*, denoted* $\mathcal{B}_{\mathbf{v}}$ *if and only if* $\sum_{d=1}^{r} \lambda_d u_{ds+k} = 0$ *for* $k = 1, \ldots, s$*.*

As shown in [15] membership in the band kernel requires that s linear forms vanish; the probability of this occurrence is q^{-s}. They then show that given two maps in the same band kernel, the probability that they are in the kernel of the same band space map is q^{-1}. Therefore the complexity of searching for a second vector given one vector in a band kernel is q^{s+1}. Since $\mathbf{A}$ is singular with probability approximately q^{-1} for sufficiently large q, the total probability of randomly selecting two vectors that are simultaneously in the kernel of the same band space map is q^{-s-2}.

While in [15] it was noted that there are some dependencies in the linear systems resulting in the need to search through a nontrivial space in the case that the characteristic is 2 or 3, it was discovered in [17] that we can add constraints to the system reducing the dimension and eliminating the search. Therefore the complexity of searching for a band space map is the same for all fields. The techniques in [17] can also be adapted to require only 2 band space maps for key recovery, the second of which can be found more cheaply by reusing one of the vectors used to find the first band space map. Since we have to compute the rank of an $n \times n$ matrix for each guess, the complexity of the attack is $\mathcal{O}(n^{\omega} q^{s+2})$ including the linear algebra overhead.

4 Combinatorial Key Recovery, the Rectangular Case

The change from square instances of the Simple Matrix scheme to rectangular instances was proposed in [30] as a way of improving efficiency by having smaller fields while maintaining a low decryption failure rate. Still requiring a left inverse of $\mathbf{A}$, the proposal requires that $r > s$. Notice, however, that this implies that there is a nontrivial left kernel of $A(\mathbf{x})$ for any vector $\mathbf{x}$!

Specifically, notice that since there are more rows than columns in $\mathbf{A}$ for the new parameters, there is always a linear combination of the rows producing the zero vector for any input. Thus, there is no search through plaintexts to find a vector in some band kernel.

In fact, the situation is worse. Note that any plaintext $\mathbf{x}$ is guaranteed to produce an $\mathbf{A}$ for which there are $r - s$ linearly independent combinations of row vectors producing zero. Therefore $\mathbf{x}$ is in very many distinct band spaces. This fact reduces the complexity of finding a second vector in the band kernel considerably, as we now show.

4.1 The Probability of Choosing a Second Band Kernel Vector

A vector $\mathbf{u} = (u_1, u_2, \ldots, u_{rs})$ belongs to a band kernel $\mathcal{B}_\mathbf{v}$ if there is a nonzero vector $\mathbf{v} \in \mathbb{F}^r$ such that for $i = 1, \ldots, s$

$$\mathbf{v} \cdot \mathbf{u}_i = 0, \text{ where } \mathbf{u}_i = (u_i, u_{r+i}, \ldots, u_{((r-1)s+i)}).$$

That is, each subvector $\mathbf{u}_i$ belongs to the orthogonal space $\langle \mathbf{v} \rangle^\perp$.

Since the space $\langle \mathbf{v} \rangle^\perp$ has dimension $r - 1$, membership of each subvector in this space can be modeled as the satisfaction of one linear relation; therefore, there are a total of s linear constraints on $\mathbf{u}$ defining membership in the $\mathcal{B}_\mathbf{v}$. Thus, for any uniformly chosen vector $\mathbf{u} \in \mathbb{F}^{rs}$ we have

$$Pr\left(\mathbf{u} \in \mathcal{B}_\mathbf{v}\right) = q^{-s}.$$

Now consider a vector $\mathbf{w} \in \mathbb{F}^r$ linearly independent with $\mathbf{v}$. The dimension of the orthogonal space $(\mathbf{w} \oplus \mathbf{v})^\perp$ is $r - 2$. Thus by the same reasoning as above,

$$Pr\left(\mathbf{u} \in \mathcal{B}_\mathbf{w} \cap \mathcal{B}_\mathbf{v}\right) = q^{-2s}.$$

In the case $r = s + 1$, we are assured that a plaintext $\mathbf{x}$ gives us $\mathbf{u} \in \mathcal{B}_\mathbf{v}$, where $\mathbf{u} = U\mathbf{x}$. Therefore membership of a second vector in the same band kernel occurs with probability q^{-s}, and the complexity of finding the second vector is q^s.

In the case that $r > s + 1$, for each plaintext $\mathbf{x}$ we are guaranteed that there are $r - s$ linearly independent vectors $\mathbf{v}_1, \ldots, \mathbf{v}_{r-s}$ such that $\mathbf{u} \in \mathcal{B}_{\mathbf{v}_i}$. Therefore $\mathbf{u}$ belongs to

$$\ell = \frac{q^{r-s} - 1}{q - 1} = q^{r-s-1} + q^{r-s-2} + \cdots + q + 1$$

distinct band kernels. Let them be $\mathcal{B}_{v_1}, \mathcal{B}_{v_2}, \ldots, \mathcal{B}_{v_\ell}$. Here it might be the case that $\mathcal{B}_{v_i} \cap \mathcal{B}_{v_j} \neq \mathcal{B}_{v_s} \cap \mathcal{B}_{v_k}$, but all the intersections have the same dimension $rs - 2s$. So, the probability $\mathbf{u}$, chosen at random, belongs to one of them is roughly

$$\begin{aligned} Pr\left(\mathbf{u} \in \bigcup_{k=1}^{\ell} \mathcal{B}_{v_k}\right) &\approx \frac{(\sum_{i=0}^{r-s-1} q^i) q^{rs-s} - \binom{\sum_{i=0}^{r-s-1} q^i}{2} q^{rs-2s}}{q^{rs}} \\ &\approx q^{r-2s-1} - q^{2(r-2s-1)} \\ &\approx q^{r-2s-1}. \end{aligned}$$

Thus, the complexity of finding a second band kernel vector is roughly q^{2s+1-r}.

4.2 The effect of $u + v > 2s$

A further effect of the rectangular augmentation of the Simple Matrix Scheme is that it requires the number of columns of the matrices $\mathbf{B}$ and $\mathbf{C}$ to be increased for efficiency. We therefore find that all of the proposed parameters with $q < 2^{32}$ have $u + v \geq 2s + 4$.

Theorem 1 *If $\mathbf{x}_1$ and $\mathbf{x}_2$ fall within the band kernel $\mathcal{B}_{\mathbf{v}}$, then they are both in the kernel of some generalized band-space differential $\mathbf{DQ} = \sum_{Q_i \in \mathcal{B}_{\mathbf{v}}} \tau_i \mathbf{DQ}_i$ with probability approximately q^{-1} if $u + v = 2s$ and probability 1 if $u + v > 2s$. Further, if $u + v > 2s$ then there exists, with probability 1, some $(u + v - 2s)$-dimensional subspace of $\mathcal{B}_{\mathbf{v}}$, all elements of which have both vectors in their kernel.*

Proof. There are two cases: (i) $u + v = 2s$ and (ii) $u + v > 2s$. The first case follows exactly from [15, Theorem 2]. The second case is new, so we focus on this second case in what follows. This will be quite similar to the original proof, but we include the full details for the reader.

A $\mathbf{DQ}$ meeting the above condition exists iff there is a nontrivial solution to the following system of equations

$$\begin{aligned} \sum_{Q_i \in \mathcal{B}_v} \tau_i \mathbf{DQ}_i \mathbf{x_1}^T &= 0 \\ \sum_{Q_i \in \mathcal{B}_v} \tau_i \mathbf{DQ}_i \mathbf{x_2}^T &= 0. \end{aligned} \tag{1}$$

Expressed in a basis where the first s basis vectors are chosen to be outside the band kernel, and the remaining $n - s$ basis vectors are chosen from within the band kernel, the band-space differentials take the form:

$$\mathbf{DQ}_i = \left[\begin{array}{c|c} \mathbf{S}_i & \mathbf{R}_i \\ \hline \mathbf{R}_i^T & 0 \end{array}\right], \tag{2}$$

where R_i is a random $s \times n - s$ matrix and S_i is a random symmetric $s \times s$ matrix. Likewise $\mathbf{x_1}$ and $\mathbf{x_2}$ take the form $(0|\ \mathbf{x_k}\ )$. Thus removing the redundant degrees of freedom we have the system of $2s$ equations in $u + v$ variables:

$$\begin{aligned} \sum_{i=1}^{u+v} \tau_i \mathbf{R}_i {\mathbf{x_1}}^T &= 0 \\ \sum_{i=1}^{u+v} \tau_i \mathbf{R}_i {\mathbf{x_2}}^T &= 0. \end{aligned} \tag{3}$$

This has a nontrivial solution precisely when the following matrix has a nontrivial right kernel:

$$\left[\begin{array}{c|c|c|c} | & | & & | \\ \mathbf{R}_1{\mathbf{x_1}}^T & \mathbf{R}_2{\mathbf{x_1}}^T & \cdots & \mathbf{R}_{u+v}{\mathbf{x_1}}^T \\ | & | & & | \\ \hline | & | & & | \\ \mathbf{R}_1{\mathbf{x_2}}^T & \mathbf{R}_2{\mathbf{x_2}}^T & \cdots & \mathbf{R}_{u+v}{\mathbf{x_2}}^T \\ | & | & & | \end{array}\right] \tag{4}$$

By the assumption that $u + v > 2s$, this matrix has more columns than rows, and therefore must have a nontrivial right kernel with probability 1. Moreover, with probability 1, this right kernel has dimension at least $u + v - 2s$. Therefore, any differential produced by taking the direct product of $(Q_1, ..., Q_{u+v})$, where $Q_1, ..., Q_{u+v}$ are the generators of $\mathcal{B}_v$, and a right kernel vector of the aforementioned matrix will have both $\mathbf{x}_1$ and $\mathbf{x}_2$ in its kernel.

4.3 Controlling the Ratio $\frac{m}{n}$.

The new parameters presented in [30] added another feature to Simple Matrix: the ability to decouple the number of variables n from the size rs of the matrix $\mathbf{A}$. The authors want to ensure that the number of equations is not significantly more than twice the number of variables so that the first fall degree of the system is not diminished.

In all but the case of $q = 2^8$, the authors of [30] propose parameters with $m = 2n$. In the case of $q = 2^8$, however, the relationship is more complicated. All parameters in this case are functions of s. Specifically, $r = s + 3$, $u = v = s + 4$, $n = s(s + 8)$ and $m = 2(s + 3)(s + 4)$. Therefore $m - 2n = 24 - 2s$.

For small s, this change poses a challenge to the linear algebra search MinRank method. The reason is that choosing merely two kernel vectors results in a system that is underdetermined, and since the size of the field is still fairly large $q = 2^8$, it is very costly to search through the solution space. On the other hand, if we increase the number of kernel vectors we guess to three, we have an additional factor of q^s in our complexity estimate.

Luckily, there is an easy way to handle this issue. We simply ignore some of the public differentials. Consider the effect of removing a of the public equations on the MinRank attack. If $a \geq m - 2n$, then we need to only consider $\lceil \frac{m-a}{n} \rceil = 2$ kernel vectors. Since the expected dimension of the intersection of any band space, which is of dimension $u + v$, and the span of the $m - a$ remaining public maps is $(u+v)+(m-a)-m = u+v-a$, we can apply Theorem 1 with $u+v-a$ in place of $u + v$. We have now shown the following:

Corollary 1 *Consider the public key P with a equations removed. Let $\widehat{\mathcal{B}}_{\mathbf{v}}$ be the intersection of the band space $\mathcal{B}_{\mathbf{v}}$ and the remaining public maps. If $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}_{\mathbf{v}}$, then there exists a band-space differential $\mathbf{DQ} = \sum_{Q_i \in \widehat{\mathcal{B}}_{\mathbf{v}}} \tau_i \mathbf{DQ}_i$ whose kernel contains both $\mathbf{x}_1$ and $\mathbf{x}_2$ with probability approximately q^{-1} if $u+v-a = 2s$ and probability 1 if $u+v-a > 2s$. Further, if $u+v-a > 2s$ then there exists, with probability 1, some $(u+v-a-2s)$-dimensional subspace of $\widehat{\mathcal{B}}_{\mathbf{v}}$, all elements of which have both vectors in their kernel.*

Considering the parameters from [30], we see that the largest value of a required to produce a fully determined MinRank system with two kernel vectors is in the case that $s = 8$ producing $a = m - 2n = 24 - 2s = 8$. In this same set of parameters $u = v = s + 4$ so that $u + v = 2s + 8$. Therefore, Corollary 1 applies.

5 Improvements from a Reaction Attack

As noted in [30], the original proposal of the square version of Simple Matrix, cf. [29], did not properly address decryption failures. To maintain performance and avoid the attack from [15], the rectangular scheme was introduced with many possible field sizes. Still, the proposed parameters only made decryption failures less common but not essentially impossible. The smallest decryption failure rate for parameters in [30] is 2^{-64} and the only parameters with sufficiently good performance to advertise had decryption failure rates of 2^{-32}.

These augmentations addressed decryption failures out of precaution, but had no claim of how such failures could be used to undermine the scheme. In this section we develop an enhancement of our combinatorial key recovery from the previous section utilizing these decryption failures. To our knowledge, this is the first example of a key recovery reaction attack against a multivariate scheme in this context.

5.1 Decryption Failures in the Simple Matrix scheme

As described in Subsection 2.2, the decryption algorithm of Simple Matrix assumes that the matrix $A(\mathbf{u})$ has a left inverse. This property is exactly the same assumption in the more general case of rectangular matrices as well. The failure of $A(\mathbf{u})$ to be full rank makes the decryption algorithm fail, producing decryption failures. One could imagine guessing which rows of $\mathbf{WA}$ could be made into elementary basis vectors trying to recover linear relations on the values of $\mathbf{u}$ to recover a quadratic system in fewer variables which may produce an unique preimage, but this is costly in performance and still allows an adversary to detect when $A(\mathbf{u})$ is not of full rank.

If we consider $\mathbf{A}$ to be rectangular, say $r \times s$, then we need the number of rows r to be greater than or equal to s. Then we may still have a left inverse $\mathbf{W}$, an $s \times r$ matrix satisfying $\mathbf{WA} = \mathbf{I}_s$. The probability of the existence of at least one such $\mathbf{W}$ is the same as the probability that the rows of $\mathbf{A}$ span $\mathbb{F}^s$. Thus

$$Pr(\text{Rank}(\mathbf{A}) < s) = 1 - \prod_{i=r-s+1}^{r} (1 - q^{-i}) \approx q^{s-r-1}.$$

Notice that decryption failure reveals precise information about the internal state of the decryption algorithm. Specifically, the quantity $A(\mathbf{u})$ where $\mathbf{u} = U(\mathbf{x})$ is not of full rank. Even for very large q, one requires a disparity in the values r and s to make the decryption failure rate very low. Even for the parameters proposed having the smallest decryption failure rate, $q = 2^{32}$ and $r = s + 1$, the probability of decryption failure is 2^{-64} and 2^{64} decryption queries on average are needed to detect a decryption failure.

5.2 The Reaction Attack

Consider, for a moment, the square case of the Simple Matrix Scheme, that is, when $r = s$. In the search process, you try to find two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ that are simultaneously in the kernel of the same linear combination of the public differentials. For the search to succeed in finding a band map you need three events to simultaneously occur: (P_1) $\mathbf{x}_i$ to be in the band kernel of a band space; (P_2) $\mathbf{x}_{3-i}$ to be in the band kernel of the same band space; and, (P_3) for them to both be in the kernel of the same band space map.

The probability of these events occurring simultaneously is then

$$Pr(P_1 \wedge P_2 \wedge P_3) = Pr(P_1)Pr(P_2|P_1)Pr(P_3|P_1 \wedge P_2) = q^{-1} \cdot q^{-s} \cdot q^{-1} = q^{-s-2}.$$

So, it takes q^{s+2} guesses in expectation to succeed in finding two such vectors and thereby recover a band space.

Notice that decryption failure occurs when the matrix $\mathbf{A}$ is singular, which is exactly the condition for membership in some band kernel. Thus, the first vector lying in a band kernel need not be found by search. If you already have access to a decryption failure producing plaintext, then the first condition is satisfied saving a factor of q in complexity at the cost of q decryption queries. So this component of q is now, in some sense, additive instead of multiplicative in the complexity analysis of the attack. Therefore, if the decryption failure rate is sufficiently low, a reaction attack can be employed.

We find that decryption failures provide a similar advantage in the rectangular case as well. When $r > s$, a decryption failure $\mathbf{x}$ assures the existence of $r - s + 1$ linearly independent vectors $v_1, \ldots, v_{r-s+1} \in \mathbb{F}^r$ such that $\mathbf{u} \in \mathcal{B}_{v_1} \cap \cdots \cap \mathcal{B}_{v_{r-s+1}}$, where $\mathbf{u} = U\mathbf{x}$. Thus we know for sure there are

$$\ell = \frac{q^{r-s+1} - 1}{q - 1} = q^{r-s} + q^{r-s-1} + \cdots + q + 1$$

distinct band kernel spaces in which $\mathbf{u}$ belongs. Let them be $\mathcal{B}_{v_1}, \mathcal{B}_{v_2}, \dots, \mathcal{B}_{v_\ell}$. Here it might be the case that $\mathcal{B}_{v_i} \cap \mathcal{B}_{v_j} \neq \mathcal{B}_{v_s} \cap \mathcal{B}_{v_k}$, but all the intersection have the same dimension $rs - 2s$. So, the probability that $\mathbf{u}$, chosen at random, belongs to one of them is roughly

$$\begin{aligned} Pr\left(\mathbf{u} \in \bigcup_{k=1}^{\ell} \mathcal{B}_{v_k}\right) &\approx \frac{(\sum_{i=0}^{r-s} q^i) q^{rs-s} - \binom{\sum_{i=0}^{r-s} q^i}{2} q^{rs-2s}}{q^{rs}} \\ &\approx q^{r-2s} - q^{2(r-2s)} \\ &\approx q^{r-2s}. \end{aligned}$$

6 Complexity

Noting that there is statistically no difference between using the input transformation U and choosing $\mathbf{A}$ to consist of random linear forms, we note that full key extraction including the input and output transformations proceeds as in [17]. Since this last part occurs after the recovery of the band spaces, it is of additive complexity. Therefore the complexity of the attack is equivalent to the MinRank step plus some additive overhead.

Recovering a band space then requires q^{2s+1-r} iterations of solving a linear system of size n and rank calculations on a matrix of size n. (Note that in this case, finding the second map from the same band space is cheaper by a factor of q.) Thus the complexity of the combinatorial key recovery is $\mathcal{O}(n^\omega q^{2s+1-r})$, where ω is the linear algebra constant. We note that in practice that assuming ω takes a value of approximately 2.8 results in a big-oh constant of less than one.

In the case of the reaction attack, recovering two maps from a band space requires only $2q^{2s-r}$ iterations of system solving and rank calculations. Therefore, for the reaction attack, the complexity is $\mathcal{O}(n^\omega q^{2s-r})$. The actual complexity in field operations for completing the attacks are listed in Table 3.

Using SAGE[1] [31], we performed some minrank computations on small scale variants of the ABC scheme. The computations were done on a computer with a 64 bit quad-core Intel i7 processor, with clock cycle 2.8 GHz. We were interested in verifying our complexity estimates on the most costly step in the attack, the MinRank instance, rather than the full attack on the scheme. Given as input the finite field size q, and the scheme parameter s, we computed the average number of vectors $\mathbf{x}$ required to be sampled in order to recover a matrix of rank $2s$. For our first experiment we set our parameters to $u = v = s + 1$, $r = s + 2$, and $n = ru = (s + 1)(s + 2)$. Our results are provided in Table 1.

For higher values of q and s the computations took too long to produce sufficiently many data points and obtain meaningful results with SAGE. Our analysis predicted the number of vectors needed would be on the order of Exp=$(q - 1)q^{s-3}$. Table 1 shows the comparison between our experiments and the expected value. We only used a small number of trials, particularly for the higher values of s listed for each q.

[1] Any mention of commercial products does not indicate endorsement by NIST.

We also ran another experiment exhibiting the behavior of the attack when $2n > m$. We used $u = v = s + 1$, $r = s + 2$, and $n = ru - 1 = s^2 + 3s + 1$. We then threw away two of the equations generated. Our analysis predicted the number of trials required to be roughly $(q-1)q^{s-2}$. The resulting data are given in Table 2. The expected number of trials was Exp=$(q-1)q^{s-2}$.

		$s = 3$	Exp	$s = 4$	Exp	$s = 5$	Exp	$s = 6$	Exp	$s = 7$	Exp
	$q = 2$	1.8	1	3.0	2	4.7	4	6.6	8	15.8	16
	$q = 3$	2.5	2	6.6	6	19.1	18	53.1	54	173	162
	$q = 4$	3.1	3	11.9	12	47.6	48	189.0	192		
!htbp	$q = 5$	4.3	4	20.6	20	99.4	100	520.8	500		
	$q = 7$	6.5	6	40.6	42	281.4	284	1873	1988		
	$q = 8$	8	7	62.8	56	444.2	448				
	$q = 11$	9.8	10	113.6	110	1318.8	1210				
	$q = 13$	11.7	12	157.7	156	2026.7	2028				

Table 1. Average number of vectors needed for the rank to fall to $2s$. This experiment used $u = v = s + 1$, $r = s + 2$, and $n = ru = (s + 1)(s + 2)$.

		$s = 3$	Exp	$s = 4$	Exp	$s = 5$	Exp	$s = 6$	Exp
	$q = 2$	1.9	2	4.4	4	7	8	14.5	16
	$q = 3$	5.9	6	16.3	18	47	54	138.9	162
!htbp	$q = 5$	17.9	20	86.2	100	500.3	500	2137.3	2500
	$q = 7$	36.6	35	277.7	245	2092.3	1715		
	$q = 11$	100.4	110	1175	1210				
	$q = 13$	148.3	156	1855.4	2028				

Table 2. Average number of vectors needed for the rank to fall to $2s$. This experiment used $u = v = s + 1$, $r = s + 2$, and $n = ru - 1 = s^2 + 3s + 1$, and did not use two of the equations generated.

7 Conclusion

The rectangular version of the Simple Matrix Encryption Scheme is needed to avoid a high decryption failure rate and known attacks. From the analysis made in this paper, we conclude that the security of this version is actually worse than that of the square version. Furthermore, we showed that decryption failures are actually still exploitable in a concrete reaction attack that clearly undermines the security claims of the scheme.

It is interesting to consider the historical difficulty of achieving secure multivariate public key encryption. Even using the relatively new approach of defining public keys with vastly larger codomains— a change which on the surface would

Scheme	Sec. Level	a	Comb. Att. Comp.	React. Att. Comp.
$\text{ABC}(2^8, 11, 8, 12, 12, 264, 128)$	80	8	$2^{\mathbf{75.6}}$	$2^{\mathbf{67.6}}$
$\text{ABC}(2^8, 12, 9, 13, 13, 312, 153)$	90	6	$2^{\mathbf{76.3}}$	$2^{\mathbf{68.3}}$
$\text{ABC}(2^8, 13, 10, 14, 14, 364, 180)$	100	4	$2^{\mathbf{85.0}}$	$2^{\mathbf{77.0}}$

Table 3. Complexity of our Combinatorial MinRank and Reaction attacks against $q = 2^8$ parameters of the ABC Simple Matrix Encryption Scheme.

seem to allow much greater freedom in selecting secure injective functions— we observe that essentially none of the recent such proposals have attained their claimed level of security after scrutiny. Perhaps there is a fundamental barrier ensuring that any efficiently invertible function must have some exploitable property, such as low rank, preventing the advantage of privileged information of the legitimate user from dramatically separating the complexity of that efficient inversion from the adversary's task. It seems that multivariate encryption is an area still in need of significant development.

References

1. Daniel Cabarcas, Daniel Smith-Tone, and Javier A. Verbel. Key recovery attack for ZHFE. In Lange and Takagi [13], pages 289–308.
2. Ryann Cartor and Daniel Smith-Tone. An updated security analysis of PFLASH. In Lange and Takagi [13], pages 241–254.
3. Ryann Cartor and Daniel Smith-Tone. EFLASH: A new multivariate encryption scheme. In Carlos Cid and Michael J. Jacobson Jr., editors, *Selected Areas in Cryptography - SAC 2018 - 25th International Conference, Calgary, AB, Canada, August 15-17, 2018, Revised Selected Papers*, volume 11349 of *Lecture Notes in Computer Science*, pages 281–299. Springer, 2018.
4. M.-S. Chen, B.-Y. Yang, and D. Smith-Tone. Pflash - secure asymmetric signatures on smart cards. Lightweight Cryptography Workshop 2015, 2015. http://csrc.nist.gov/groups/ST/lwc-workshop2015/papers/session3-smith-tone-paper.pdf.
5. J. Ding and D. Schmidt. Rainbow, a new multivariable polynomial signature scheme. *ACNS 2005, LNCS*, 3531:164–175, 2005.
6. Jintai Ding. A new variant of the matsumoto-imai cryptosystem through perturbation. In Feng Bao, Robert H. Deng, and Jianying Zhou, editors, *Public Key Cryptography - PKC 2004, 7th International Workshop on Theory and Practice in Public Key Cryptography, Singapore, March 1-4, 2004*, volume 2947 of *Lecture Notes in Computer Science*, pages 305–318. Springer, 2004.
7. Jintai Ding, Albrecht Petzoldt, and Lih-chung Wang. The cubic simple matrix encryption scheme. In Mosca [18], pages 76–87.
8. Vivien Dubois, Pierre-Alain Fouque, and Jacques Stern. Cryptanalysis of SFLASH with Slightly Modified Parameters. In Moni Naor, editor, *EUROCRYPT*, volume 4515 of *Lecture Notes in Computer Science*, pages 264–275. Springer, 2007.
9. J. C. Faugere. Algebraic cryptanalysis of hidden field equations (HFE) using grobner bases. *CRYPTO 2003, LNCS*, 2729:44–60, 2003.
10. Pierre-Alain Fouque, Louis Granboulan, and Jacques Stern. Differential cryptanalysis for multivariate schemes. In Ronald Cramer, editor, *Advances in Cryptology*

- *EUROCRYPT 2005, 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005, Proceedings*, volume 3494 of *Lecture Notes in Computer Science*, pages 341–353. Springer, 2005.

11. Yasuhiko Ikematsu, Ray A. Perlner, Daniel Smith-Tone, Tsuyoshi Takagi, and Jeremy Vates. HFERP - A new multivariate encryption scheme. In Tanja Lange and Rainer Steinwandt, editors, *Post-Quantum Cryptography - 9th International Conference, PQCrypto 2018, Fort Lauderdale, FL, USA, April 9-11, 2018, Proceedings*, volume 10786 of *Lecture Notes in Computer Science*, pages 396–416. Springer, 2018.
12. A. Kipnis, J. Patarin, and L. Goubin. Unbalanced oil and vinegar signature schemes. *EUROCRYPT 1999. LNCS*, 1592:206–222, 1999.
13. Tanja Lange and Tsuyoshi Takagi, editors. *Post-Quantum Cryptography - 8th International Workshop, PQCrypto 2017, Utrecht, The Netherlands, June 26-28, 2017, Proceedings*, volume 10346 of *Lecture Notes in Computer Science*. Springer, 2017.
14. Tsutomu Matsumoto and Hideki Imai. Public Quadratic Polynominal-Tuples for Efficient Signature-Verification and Message-Encryption. In *EUROCRYPT*, pages 419–453, 1988.
15. Dustin Moody, Ray A. Perlner, and Daniel Smith-Tone. An asymptotically optimal structural attack on the ABC multivariate encryption scheme. In Mosca [18], pages 180–196.
16. Dustin Moody, Ray A. Perlner, and Daniel Smith-Tone. Key recovery attack on the cubic ABC simple matrix multivariate encryption scheme. In Roberto Avanzi and Howard M. Heys, editors, *Selected Areas in Cryptography - SAC 2016 - 23rd International Conference, St. John's, NL, Canada, August 10-12, 2016, Revised Selected Papers*, volume 10532 of *Lecture Notes in Computer Science*, pages 543–558. Springer, 2016.
17. Dustin Moody, Ray A. Perlner, and Daniel Smith-Tone. Improved attacks for characteristic-2 parameters of the cubic ABC simple matrix encryption scheme. In Lange and Takagi [13], pages 255–271.
18. Michele Mosca, editor. *Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings*, volume 8772 of *Lecture Notes in Computer Science*. Springer, 2014.
19. J. PATARIN. The oil and vinegar signature scheme. *Dagstuhl Workshop on Cryptography September, 1997*, 1997.
20. J. Patarin, N. Courtois, and L. Goubin. Flash, a fast multivariate signature algorithm. *CT-RSA 2001, LNCS*, 2020:297–307, 2001.
21. Jacques Patarin. Cryptoanalysis of the Matsumoto and Imai Public Key Scheme of Eurocrypt '88. In Don Coppersmith, editor, *CRYPTO*, volume 963 of *Lecture Notes in Computer Science*, pages 248–261. Springer, 1995.
22. Jacques Patarin. Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In *EUROCRYPT*, pages 33–48, 1996.
23. Jacques Patarin, Nicolas Courtois, and Louis Goubin. Quartz, 128-bit long digital signatures. In David Naccache, editor, *CT-RSA*, volume 2020 of *Lecture Notes in Computer Science*, pages 282–297. Springer, 2001.
24. Ray A. Perlner, Albrecht Petzoldt, and Daniel Smith-Tone. Total break of the SRP encryption scheme. In Carlisle Adams and Jan Camenisch, editors, *Selected Areas in Cryptography - SAC 2017 - 24th International Conference, Ottawa, ON,*

Canada, August 16-18, 2017, Revised Selected Papers, volume 10719 of *Lecture Notes in Computer Science*, pages 355–373. Springer, 2017.

25. A. Shamir and A. Kipnis. Cryptanalysis of the oil & vinegar signature scheme. *CRYPTO 1998. LNCS*, 1462:257–266, 1998.
26. P. W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM J. Sci. Stat. Comp.*, 26, 1484, 1997.
27. Daniel Smith-Tone and Javier Verbel. A key recovery attack for the extension field cancellation encryption scheme. in concurrent submission to PQCrypto 2020, 2020.
28. Alan Szepieniec, Jintai Ding, and Bart Preneel. Extension field cancellation: A new central trapdoor for multivariate quadratic systems. In Tsuyoshi Takagi, editor, *Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016, Proceedings*, volume 9606 of *Lecture Notes in Computer Science*, pages 182–196. Springer, 2016.
29. Chengdong Tao, Adama Diene, Shaohua Tang, and Jintai Ding. Simple matrix scheme for encryption. In Philippe Gaborit, editor, *PQCrypto*, volume 7932 of *Lecture Notes in Computer Science*, pages 231–242. Springer, 2013.
30. Chengdong Tao, Hong Xiang, Albrecht Petzoldt, and Jintai Ding. Simple matrix - A multivariate public key cryptosystem (MPKC) for encryption. *Finite Fields and Their Applications*, 35:352–368, 2015.
31. The Sage Developers. *SageMath, the Sage Mathematics Software System (Version 8.7)*, 2019. `https://www.sagemath.org`.
32. Christopher Wolf, An Braeken, and Bart Preneel. On the security of stepwise triangular systems. *Des. Codes Cryptogr.*, 40(3):285–302, 2006.
33. Bo-Yin Yang and Jiun-Ming Chen. Building secure tame-like multivariate public-key cryptosystems: The new TTS. In Colin Boyd and Juan Manuel González Nieto, editors, *Information Security and Privacy, 10th Australasian Conference, ACISP 2005, Brisbane, Australia, July 4-6, 2005, Proceedings*, volume 3574 of *Lecture Notes in Computer Science*, pages 518–531. Springer, 2005.
34. Takanori Yasuda and Kouichi Sakurai. A multivariate encryption scheme with rainbow. In Sihan Qing, Eiji Okamoto, Kwangjo Kim, and Dongmei Liu, editors, *Information and Communications Security - 17th International Conference, ICICS 2015, Beijing, China, December 9-11, 2015, Revised Selected Papers*, volume 9543 of *Lecture Notes in Computer Science*, pages 236–251. Springer, 2015.

Poster: Method for Effective Measurement, Labeling, and Classification of Botnet C2s for Predicting Attacks

Mitsuhiro Hatada
National Institute of Standards and Technology
mitsuhiro.hatada@nist.gov

Matthew Scholl
National Institute of Standards and Technology
matthew.scholl@nist.gov

***Abstract*—In the era of the Internet of Things, botnet threats are rising, which has prompted many studies on botnet detection. This study aims to detect the early signs of botnet attacks such as massive spam emails and Distributed Denial-of-Service attacks. To that end, this study develops a practical method for measurement, labeling, and classification of botnet Command and Control (C2) for predicting attacks. The focus is on C2 traffic and measurement of the comprehensive metrics studied in previous works. The data is labeled based on the result of the correlation analysis between C2 metrics and spam volume. Then, a special type of recurrent neural network, i.e., Long Short-Term Memory, is applied to detect an increase in spam by a botnet. The proposed method managed to detect it with an accuracy of 0.981.**

I. Introduction

A botnet is still a serious threat to cybersecurity as it controls a massive number of compromised hosts to conduct various attacks such as sending email spam or launching a Distributed Denial-of-Service (DDoS) attack. A Command and Control (C2) server plays a significant role in a botnet: it sends commands to bots and receives outputs of bots while hiding a botmaster behind it. Despite many previous attempts at botnet measurement [6], [8] and botnet detection [4], [5], [7], [10], to the best of our knowledge, there is no study that *detects the early signs of botnet attacks*. According to the botnet communication patterns [11], once a bot infection occurs, the bot registers itself to C2 and the keep-alive communication between the bot and C2 starts periodically. A botmaster issues a command to bots for launching an attack through C2, and then an attack is launched. The key idea in this study is that traffic patterns from or to C2 will be changed before the attack. For example, a botmaster would prefer more bots to launch an attack effectively, and then the bot register may increase before the attack. A command contains some data such as the target and parameter, and will be sent to numerous bots in parallel so that the size of the packet may be increased before the attack. Such predictive threat intelligence is crucial for an internet service provider (ISP) for prioritizing C2s and cutting off communication between C2 and bots in advance in crisis situations or at customers' requests.

This poster presents a method for effective measurement, labeling, and classification of botnet C2s for predicting attacks. In the measurement phase, various metrics of C2 are computed with flow data collected with a certain sampling rate by an ISP for network management. Next, a set of C2 metrics is labeled for a certain period based on 1) the gradient of spam volume and 2) the result of the correlation analysis between the moving average of each C2 metric and the volume of spam email associated with C2 as attack data. For the classification, a recurrent neural network is used to train and test the labeled dataset. The following sections describe each phase of the method and preliminary experimental results.

II. Methodology

A. Measurement

The 17 metrics listed in TABLE I are defined for measuring C2 activity, which is a compilation of various metrics studied in previous works [4], [10]. All metrics are computed with flow data for every three hours. Because the used flow data has a unidirectional format, IP addresses of C2 will be observed in the source or destination field. The metrics can be computed using three patterns: C2 in source, C2 in destination, and C2 in either source or destination. Finally, because the 17 metrics are multiplied by the three patterns, it is possible to use 51 metrics. The lists of C2s are retrieved from websites [1], [3] and provided by a reliable research institution once daily.

TABLE I. C2 Metrics

Category	Metrics
Size	1) # of bots
	2) # of bots observed multiple flows
Volume	3) Average, 4) standard deviation and 5) sum of bytes
	6) Average, 7) standard deviation and 8) sum of packets
Frequency	9) # of flows
	10) # of flows with few packets (less than three packets)
	11) # of flows with short duration (less than one s)
	12) # of flows with small bytes (less than 500 bytes)
Load	13) Average, 14) standard deviation and 15) sum of duration
Lifetime	16) # of days flow was observed in the last seven days
	17) # of days flow was continuously observed in the last seven days

B. Labeling

Spam reputation data [2] is retrieved as attack data once daily. It includes information such as the IP address of the spam sender and spam volume in the last day. The total spam volume associated with C2 can be added up by associating an IP address of C2 and an IP address of the spam sender in flow data. As a preliminary experiment, it was analyzed whether the metrics could be useful for predicting the increase in spam during a week. At that time, the moving average of each C2 metric was taken with a one-day time window. This time window represents how far in advance a sign can be detected. To align the number of data elements with the

Hatada, Mitsuhiro; Scholl, Matthew. "Method for Effective Measurement, Labeling, and Classification of Botnet C2s for Predicting Attacks." Paper presented at 27th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, US. February 23, 2020 - February 26, 2020.

C2 metric, spam volume was padded with the same value as the previous data. According to the result of the correlation analysis between each C2 metric and spam volume for each botnet, a different botnet tends to show different correlation and there are metrics with a high positive correlation. Based on this observation, it was decided to use all metrics and the following two criteria were set for labeling: 1) there is at least one metric with a high correlation greater than or equal to 0.3; 2) there is a positive gradient of the spam volume. With these criteria, C2 with the increasing spam for one week can be set as a *True* label. These steps are repeated for the entire period while shifting the starting point by one day.

C. Classification

Long Short-Term Memory (LSTM) [9] was applied for binary classification. LSTM is a recurrent neural network capable of learning long-term dependencies. To apply LSTM, each C2 metric is scaled between 0 and 1, and then the time series data of each C2 metric is laterally shifted. Finally, it becomes 2,856-dimensional data. Various models have been tried, but a model of stacked LSTMs was selected because of its high accuracy. The model has a layered structure comprising LSTM, dropout, LSTM, dropout, and dense. A dropout layer is used for the regularization that randomly sets some of the dimensions of the input vector to zero at each update during training time, which helps prevent overfitting. The dense layer represents matrix vector multiplication. The values in the matrix are the trainable parameters that get updated during backpropagation. The model is configured using binary cross-entropy for the loss function, Adam as the optimization algorithm, and Sigmoid as the activation function.

III. Evaluation

A. Dataset

TABLE II describes the dataset used in the evaluation, which was generated between August 3, 2019 and November 1, 2019 (90 days) and labeled as described in Section II-B with a different time window of the moving average. The number of true and false data is unbalanced, so false data are randomly sampled to align with the number of true data for experiments.

TABLE II. Dataset

Time Window (hours)	# of True	# of False	# of C2s
12	1,893	8,641	234
24	2,291	8,154	234
48	2,423	7,888	233

B. Experiment

The proposed method was evaluated with respect to accuracy as well as computational time. Using Python 3.6.9 with Keras 2.2.4 on top of TensorFlow 1.14.0, experiments were performed for training and testing with 10-fold cross-validation on the *Enki* which is the High Performance Computing cluster at the National Institute of Standards and Technology. TABLE III demonstrates the average results of the 10-fold cross-validation. The highest accuracy is 0.981, when the time window of the moving average is 48 and the number of hidden units is 100. By increasing the number of hidden units, both training time and testing time almost linearly increased as expected. The effect of the time window is not noticeable because all data might have been properly learned.

TABLE III. Results of binary classification

(batch=128, epoch=200, dropout=0.3, learning rate=0.01)

Time Window (hours)	Units	Train (s)	Test (s)	Accuracy	Recall	Precision
12	100	2,201.90	2.37	0.923	1.000	0.883
	200	3,186.01	3.71	0.929	0.956	0.918
	300	4,256.90	5.93	0.938	1.000	0.902
	400	5,566.59	9.39	0.898	0.994	0.859
	500	7,037.26	13.69	0.934	0.999	0.890
24	100	2,658.80	2.98	0.956	0.998	0.923
	200	3,817.19	4.20	0.926	1.000	0.881
	300	4,976.25	6.98	0.921	0.998	0.882
	400	6,730.17	11.15	0.966	1.000	0.939
	500	8,535.01	16.35	0.955	0.986	0.931
48	100	2,830.37	2.97	0.981	1.000	0.966
	200	4,050.56	4.56	0.881	1.000	0.829
	300	5,354.73	7.21	0.805	0.642	0.968
	400	7,138.42	11.92	0.918	0.963	0.897
	500	9,056.58	17.29	0.939	0.906	0.973

IV. Conclusions and future work

By focusing on C2 traffic, the proposed method managed to detect an increase in spam email by a botnet with an accuracy of 0.981. The next challenge is how far in advance to predict the increase in spam email. It is necessary to extract C2 communication appropriately because C2 metrics are computed including legitimate flow if a legitimate server is compromised and used as C2. It is also essential to accurately track a C2 IP address because C2 changes the IP address to avoid detection. Although the method can be further improved, the method would also be applicable to the prediction of attacks such as DDoS if the data that associates bots with C2 is available.

References

[1] abuse.ch SSLBL Snort / Suricata Botnet C2 IP Ruleset. [Online]. Available: https://sslbl.abuse.ch/blacklist/sslipblacklist.rules

[2] Email & Spam Data. [Online]. Available: https://talosintelligence.com/reputation_center/email_rep#spam-ip-senders

[3] Master Feed of known, active and non-sinkholed C&Cs IP addresses. [Online]. Available: https://osint.bambenekconsulting.com/feeds/c2-ipmasterlist-high.txt

[4] E. Biglar Beigi, H. Hadian Jazi, N. Stakhanova, and A. A. Ghorbani, "Towards effective feature selection in machine learning-based botnet detection approaches," in *Proceedings of the 2014 IEEE Conference on Communications and Network Security*, Oct 2014, pp. 247–255.

[5] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, and C. Kruegel, "Disclosure: Detecting Botnet Command and Control Servers Through Large-scale NetFlow Analysis," in *Proceedings of the 28th Annual Computer Security Applications Conference*, 2012, pp. 129–138.

[6] W. Chang, A. Mohaisen, A. Wang, and S. Chen, "Measuring Botnets in the Wild: Some New Trends," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, 2015, pp. 645–650.

[7] G. Gu, J. Zhang, and W. Lee, "BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic," in *Proceedings of the 15th Annual Network and Distributed System Security Symposium*, 2008.

[8] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin, "Measurement and Analysis of Hajime, a Peer-to-peer IoT Botnet," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, 2019.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] G. Kirubavathi and R. Anitha, "Botnet detection via mining of traffic flow characteristics," *Computers & Electrical Engineering*, vol. 50, pp. 91–101, 2016.

[11] G. Vormayr, T. Zseby, and J. Fabini, "Botnet Communication Patterns," *IEEE Communications Surveys Tutorials*, vol. 19, no. 4, pp. 2768–2796, 2017.

Hatada, Mitsuhiro; Scholl, Matthew. "Method for Effective Measurement, Labeling, and Classification of Botnet C2s for Predicting Attacks."

National Institute of Standards and Technology
U.S. Department of Commerce

Method for Effective Measurement, Labeling, and Classification of Botnet C2s for Predicting Attacks

Mitsuhiro Hatada and Matthew Scholl, Computer Security Division

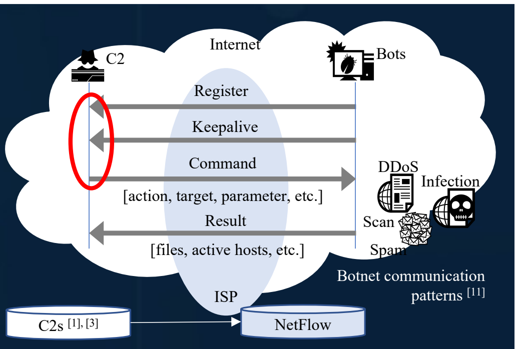
Motivation

- Botnet is still a serious threat to cybersecurity as it controls a massive number of compromised hosts to conduct various attacks such as sending email spam or launching a Distributed Denial-of-Service (DDoS) attack.
- Command and Control (C2) server plays a significant role in a botnet: it sends commands to bots and receives outputs of bots while hiding a botmaster behind it.
- Despite many previous attempts at botnet measurement [6], [8] and botnet detection [4], [5], [7], [10], to the best of our knowledge, there is no study that **detects the early signs of botnet attacks**.

Key Idea

Traffic patterns from or to C2 will be changed before the attack.

- Botmaster would prefer more bots to launch an attack effectively, and then the bot register may increase before the attack.
- Command contains some data such as the target and parameter, and will be sent to numerous bots in parallel so that the size of the packet may be increased before the attack.



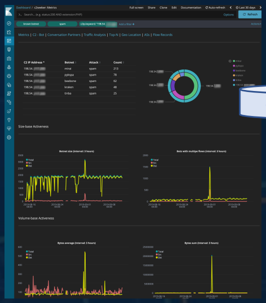
Botnet communication patterns [11]

Measurement

- All metrics are computed with flow data for every three hours and can be computed using three patterns: C2 in source field, C2 in destination field, and C2 in either source or destination field of unidirectional NetFlow.
- Finally, it is possible to use 51 metrics.

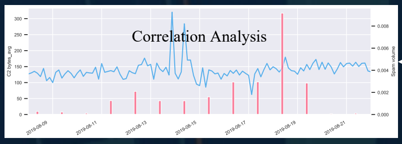
TABLE I. C2 METRICS

Category	Metrics
Size	1) # of bots
	2) # of bots observed multiple flows
Volume	3) Average, 4) standard deviation and 5) sum of bytes
	6) Average, 7) standard deviation and 8) sum of packets
Frequency	9) # of flows
	10) # of flows with few packets (less than three packets)
	11) # of flows with short duration (less than one s)
	12) # of flows with small bytes (less than 500 bytes)
Load	13) Average, 14) standard deviation and 15) sum of duration
Lifetime	16) # of days flow was observed in the last seven days
	17) # of days flow was continuously observed in the last seven days



C2 Metrics

Spam Vol. [2]



Preliminary Analysis

Could the metrics be useful for predicting the increase in spam during a week?

- The total spam volume associated with C2 can be added up by associating an IP address of C2 and an IP address of the spam sender in flow data.
- The moving average of each C2 metric was taken with a one-day time window. This time window represents how far in advance a sign can be detected.
- According to the result of the correlation analysis between each C2 metric and spam volume for each botnet, *a different botnet tends to show different correlation* and *there are metrics with a high positive correlation*.

Labeling

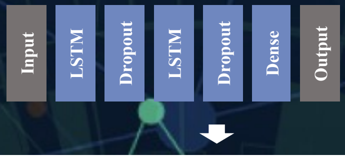
- Criteria
  1) There is at least one metric with a high correlation greater than or equal to 0.3
  2) There is a positive gradient of the spam volume
- C2 with the increasing spam for one week can be set as a True label. These steps are repeated for the entire period while shifting the starting point by one day.

C2	Time	Metric A	Metric B	…	Label	2) Gradient
198.51.100.22	8/3 00:00-03:00	133.45	1872.10	…	True	0.02
	8/3 03:00-06:00	664.87	3606.19	…		
	:	:	:	…		
	8/9 21:00-24:00	293.42	18,341.22	…		
	1) Correlation	0.11	0.42	…		

Dataset

Classification

- Long Short-Term Memory (LSTM) [9], a recurrent neural network capable of learning long-term dependencies, was applied for binary classification.
- The model is configured using binary cross-entropy for the loss function, Adam as the optimization algorithm, and Sigmoid as the activation function.
- Each C2 metric is scaled between 0 and 1, and then the time series data of each C2 metric is laterally shifted. Finally, it becomes 2,856-dimensional data.



Evaluation

- Dataset was generated between August 3, 2019 and November 1, 2019 (90 days) with a different time window of the moving average.
- The number of true and false data is unbalanced, so false data are randomly sampled to align with the number of true data for experiments.
- Experiments were performed for training and testing with 10-fold cross-validation on the Enki which is the High Performance Computing cluster at the NIST.
- By focusing on C2 traffic and using LSTM for the time series data of comprehensive metrics, the proposed method managed to detect an increase in spam email by a botnet with an accuracy of 0.981.

TABLE II. DATASET

Time Window (hours)	# of True	# of False	# of C2s
12	1,893	8,641	234
24	2,291	8,154	234
48	2,423	7,888	233

TABLE III. RESULTS OF BINARY CLASSIFICATION

(batch=128, epoch=200, dropout=0.3, learning rate=0.01)

Time Window (hours)	Units	Train (s)	Test (s)	Accuracy	Recall	Precision
12	100	2,201.90	2.37	0.923	1.000	0.883
	200	3,186.01	3.71	0.929	0.956	0.918
	300	4,256.90	5.93	0.938	1.000	0.902
	400	5,566.59	9.39	0.898	0.994	0.859
	500	7,037.26	13.69	0.934	0.999	0.890
24	100	2,658.80	2.98	0.956	0.998	0.923
	200	3,817.19	4.20	0.926	1.000	0.881
	300	4,976.25	6.98	0.921	0.998	0.882
	400	6,730.17	11.15	0.966	1.000	0.939
	500	8,535.01	16.35	0.955	0.986	0.931
48	100	2,830.37	2.97	0.981	1.000	0.966
	200	4,050.56	4.56	0.881	1.000	0.829
	300	5,354.73	7.21	0.805	0.642	0.968
	400	7,138.42	11.92	0.918	0.963	0.897
	500	9,056.58	17.29	0.939	0.906	0.973

References

[1] abuse.ch SSLBL Snort / Suricata Botnet C2 IP Ruleset. [Online]. Available: https://sslbl.abuse.ch/blacklist/sslipblacklist.rules
[2] Email & Spam Data. [Online]. Available: https://talosintelligence.com/reputation_center/email_rep#spam-ip-senders
[3] Master Feed of known, active and non-sinkholed C&Cs IP addresses. [Online]. Available: https://osint.bambenekconsulting.com/feeds/c2-ipmasterlist-high.txt
[4] E. Biglar Beigi, H. Hadian Jazi, N. Stakhanova, and A. A. Ghorbani, “Towards effective feature selection in machine learning-based botnet detection approaches,” in Proceedings of the 2014 IEEE Conference on Communications and Network Security, Oct 2014, pp. 247–255.
[5] L. Bilge, D. Balzarotti, W. Robertson, E. Kirda, and C. Kruegel, “Disclosure: Detecting Botnet Command and Control Servers Through Large-scale NetFlow Analysis,” in Proceedings of the 28th Annual Computer Security Applications Conference, 2012, pp. 129–138.
[6] W. Chang, A. Mohaisen, A. Wang, and S. Chen, “Measuring Botnets in the Wild: Some New Trends,” in Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, 2015, pp. 645–650.
[7] G. Gu, J. Zhang, and W. Lee, “BotSniffer: Detecting Botnet Command and Control Channels in Network Traffic,” in Proceedings of the 15th Annual Network and Distributed System Security Symposium, 2008.
[8] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin, “Measurement and Analysis of Hajime, a Peer-to-peer IoT Botnet,” in Proceedings of the 26th Annual Network and Distributed System Security Symposium, 2019.
[9] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[10] G. Kirubavathi and R. Anitha, “Botnet detection via mining of traffic flow characteristics,” Computers & Electrical Engineering, vol. 50, pp. 91–101, 2016.
[11] G. Vormayr, T. Zseby, and J. Fabini, “Botnet Communication Patterns,” IEEE Communications Surveys Tutorials, vol. 19, no. 4, pp. 2768–2796, 2017.

Hatada, Mitsuhiro; Scholl, Matthew. ”Method for Effective Measurement, Labeling, and Classification of Botnet C2s for Predicting Attacks.” Paper presented at 27th Annual Network and Distributed System Security Symposium (NDSS), San Diego, CA, US. February 23, 2020 - February 26, 2020.

Key Considerations for Microbial Viability Measurements

Joy P. Dunkers[1*], Sandra da Silva[1], Stephanie Servetas[1], James J. Filliben[2], Guilherme Pinheiro[3] and Nancy J. Lin[1]

[1.] Biosystems and Biomaterials Division, National Institute of Standards and Technology, Gaithersburg, MD, US
[2.] Statistical Engineering Division, NIST, Gaithersburg, MD, US
[3.] Division of Metrology Applied to Life Sciences, National Institute of Metrology, Quality and Technology, Brazil
[*]Corresponding author: joy.dunkers@nist.gov

Making reliable measurements of antimicrobial killing efficacy requires careful consideration of the sources of biological variability, measurement bias and error throughout the entire workflow. For ultraviolet-C (UV-C) disinfection, killing efficacy is most commonly measured using plate counting where the $\log_{10}$ reduction in colony forming units (CFUs) is calculated using the inoculum CFU count before and after UV-C exposure. Although CFU counting is not a rapid diagnostic, it is fit-for-purpose for testing UV-C sources and establishing room disinfection protocols. Best practices for minimizing biological and technical error in plate counting have been established [1]. Careful consideration of sources of variability during all experimental stages can help minimize error propagated to the final $\log_{10}$ reduction value. Instrument calibration, sample preparation, counting and viability measurements performed at NIST provide insights that could further improve confidence in killing efficacy results.

Samples for UV-C killing efficacy testing are prepared using bacteria at a target concentration, which is measured using a spectrophotometer and converted to bacterial concentration using an OD_{600} -bacteria concentration calibration curve. Characterization of spectrophotometer performance for linearity and offset against OD standards is an essential first step for accurate and comparable measurements. Figure 1 shows an example spectrophotometer calibration, where the linear operating range is only up to ≈ 0.9 OD, illustrating that high OD cultures should be diluted to the linear range for accurate measurement.

The OD-bacterial concentration calibration curve is typically based on CFU counts. The use of counting devices such as a Coulter (impedance) counter, flow cytometer or optical microscope can improve calibration reliability by providing an object count rather than a culturable cell count. These types of calibration curves must be collected for each strain and culture condition being studied, as the relationship between OD and concentration will vary. For instance, *E. coli* cultures diluted to the same cell concentration had very different OD readings when they were grown either in rich or in minimal media. Using alternate counting methods helps in understanding how well CFUs compare with maximum attainable count.

Microbe aggregation contributes to counting error, so disaggregation while maintaining viability can improve viable cell counts accuracy. In one approach, a focused ultrasonic instrument is used to break microbial aggregates into single cells to improve counting accuracy. *A. naeslundii*, an aggregated facultative anaerobe, is highly disaggregated after focused ultrasonic treatment optimization (Figure 2). Comparison of CFU count and Coulter counter count as a function of ultrasonic treatment time showed that an optimum treatment time can maximize viable counts of disaggregated *A. naeslundii*.

Dunkers, Joy; Da Silva, Sandra M.; Servetas, Stephanie; Filliben, James J.; Pinheiro, Guilherme; Lin, Nancy. "Key Considerations for Microbial Viability Measurements." Paper presented at Workshop on Ultraviolet Disinfection Technologies & Healthcare Associated Infections: Defining Standards and Metrology Needs, Gaithersburg, MD, US. January 14, 2020 - January 15, 2020.

The accuracy and precision of CFU measurements is also dependent on microorganism size, concentration and morphology (chain vs single). Comparison of CFUs to Coulter counter results has been performed using *S. cerevisiae* (spherical yeast), *B. thuringiensis* (spores), *B. thailandenis* (short rods), and *S. mutans* (chains). A 2 µm bead concentration standard was used for absolute count. Agreement between CFU and Coulter counter results varied by microorganism, although there was very good counting agreement between known bead concentration and Coulter counter results. *B. thailandenis* had lower precision for both CFU and Coulter counter compared to the other organisms. For *S. mutans*, plate counting had much lower precision than the Coulter counter. The addition of Coulter counter data aids in evaluating robustness in plate counting.

Orthogonal methods for viability provide independent results, albeit with different caveats. Orthogonal methods are not meant to be run regularly, but as a check on the primary method. A desirable orthogonal method to CFUs would allow for increased throughput and more robust statistics. Flow cytometry and automated optical microscopy are two high throughput methods that both rely on fluorescent probe intensity as a surrogate for viability. In selecting a fluorescent probe, the mechanism of cell death should be considered. For UV-C killing, DNA strand breakage and base rearrangement have been identified as the main killing mechanisms [2]. Membrane integrity fluorescent probes have been used as viability markers [3] but may not be suitable for testing UV-C killing efficacy. Loss of membrane potential may be considered independent of killing method, and membrane potential fluorophores as viability markers have been studied in bacteria [4] and yeast [5].

Robust measurement methods and technology innovation for cell count and viability are needed to improve confidence in plate counting as a measure of UV-C killing efficacy.

References:
[1] ASTM International. *D5465-93(2012) Standard Practice for Determining Microbial Colony Counts from Waters Analyzed by Plating Methods*. West Conshohocken, PA; ASTM International, 2012.
[2] T D Cutler and J J Zimmerman, Animal Health Res Rev **12** (2011), p. 15.
[3] M Berney, F Hammes, F Bosshard, H-U Weilenmann, T Egli, Appl Environ Microb **73(10)** (2007), p.3283.
[4] S Rezaeinejad and V Ivanov, Microb Res **166** (2011), p. 129-135.
[5] J Suchodolski and A Krasowska, Microorganisms **7** (2019), p. 110.

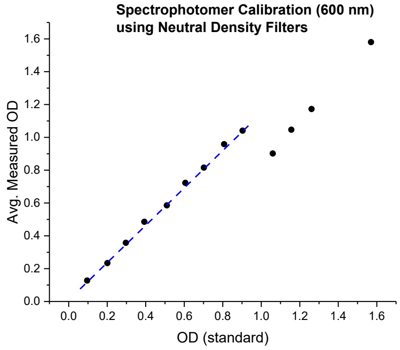


Figure 1: Linear relationship between standard and measured OD only extends to 0.9 OD.

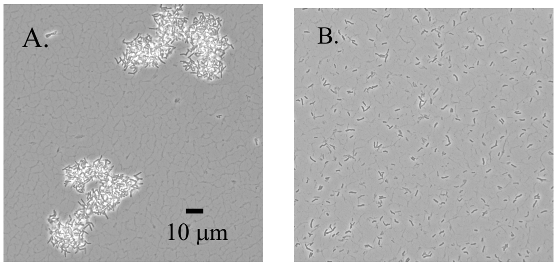


Figure 2. *A. naeslundii* before (A) and after (B) focused ultrasonic treatment.

Dunkers, Joy; Da Silva, Sandra M.; Servetas, Stephanie; Filliben, James J.; Pinheiro, Guilherme; Lin, Nancy. "Key Considerations for Microbial Viability Measurements." Paper presented at Workshop on Ultraviolet Disinfection Technologies & Healthcare Associated Infections: Defining Standards and Metrology Needs, Gaithersburg, MD, US. January 14, 2020 - January 15, 2020.

CAGEN: A fast combinatorial test generation tool with support for constraints and higher-index arrays

Michael Wagner, Kristoffer Kleine and Dimitris E. Simos
SBA Research
A-1040 Vienna, Austria
{mwagner,kkleine,dsimos}@sba-research.org

Rick Kuhn and Raghu Kacker
NIST, Information Technology Laboratory,
Gaithersburg, MD, USA
Email: {d.kuhn.nist.gov,raghu.kacker}@nist.gov

***Abstract*—In recent years, combinatorial testing methods have been successfully applied to test systems with a large number of input parameters. Generating combinatorial test sets for such complex systems is still a challenging task as it requires a lot of time and computing power. To tackle that issue, high performance tools are required. In this work, we present the combinatorial test set generation tool CAgen. It is capable of generating combinatorial test sets significantly faster than other state-of-the-art tools, such as ACTS, and contains various features such as constraint handling and higher-index arrays. It is highly compatible with other combinatorial testing tools and is available as CLI and Web-GUI. CAgen aims to make combinatorial testing more efficient and more accessible and user-friendly.**

I. INTRODUCTION

Due to the reduction in the number of tests necessary when using combinatorial testing methods, it has become possible to test larger systems within reasonable time and resources, while maintaining certain coverage guarantees. At the same time, the generation of combinatorial test sets with a small number of tests becomes increasingly difficult as the systems under test become larger [1]. To counteract this, tools are necessary that are both fast and memory-efficient. Even for smaller systems, faster generation times can lead to reduced testing cycle durations. Furthermore, in order for combinatorial testing to be adopted by a wider set of software testers and developers, the tools necessary to edit, generate and inspect combinatorial test sets need to be user-friendly, enabling a quick development cycle. Both these properties aim at lowering the entry barrier to people unfamiliar with combinatorial testing. However, presently there is a distinct gap in the space of combinatorial testing tools making a wide-spread adoption harder than necessary. In the following, we will outline a few problems of the current state-of-the-art and then propose a new tools to bridge this gap.

One of the largest problems hindering adoption is the way in which different tools are distributed. Classic software distribution via pre-compiled binaries or bundles has the drawbacks of having no easy way to implement software updates by the vendor and placing a high maintenance burden on the provider depending on the number of supported platforms. Therefore, System as a Service (SaaS) solutions are becoming a more and more popular way to bring a software product to the end user with control remaining at the side of the vendor. In case of defects, bug fixes can quickly be deployed to the service and new versions are immediately available to all users without. Recently, the authors in [2] surveyed existing combinatorial test generation tools which are available via the web. However, their analysis concludes that many of these tools have deficiencies which prevent their use in many testing scenarios. Most are commercial and thus not readily available to entry-level testers, students or academic users. Furthermore, some tools only offer limited support for higher strength test generation with three out of five tools only supporting pair-wise test generation and only one with support for 6-way coverage. As a response, in the same paper, they proposed a new tool named CTWEDGE which offers free test set generation with either ACTS or CASA accessible via a web frontend. Furthermore, in [3] a tool for automated combinatorial test generation and fault characterization was intruduced, where the algorithm underlying CAgen was used for initial test set generation.

In this work, we present CAgen[1], a fast, efficient and user-friendly combinatorial test generation tool. To evaluate its features and performance, we will draw comparisons with the following tools: for GUI functionality, we consider ACTS [4], which comes as a command line tool as well as a GUI version, and the publicly available service CTWedge [2] that offers free test set generation using ACTS or CASA. Further, we will analyze the performance of our tool against the state of the art test set generation tools ACTS and PICT.

The paper is structured as follows. Section II gives an overview about algorithms used in test set generation and constraint handling. The architecture of CAgen is explained in detail in Section III and Section IV introduces the various features of CAgen and compares them with other generation tools. Section V provides benchmarks for various different instances of covering arrays, including models from real world applications, and compares them with ACTS and PICT, while Section VI concludes our work and discusses potential future work.

II. BACKGROUND AND RELATED WORK

The efficient generation of *covering arrays* (CAs), the mathematical construct underlying combinatorial test sets, is a highly researched topic. Various generation methods, such

[1] Publicly available at https://matris.sba-research.org/tools/cagen

Kacker, Raghu N.; Kuhn, D. Richard. "CAgen: A fast combinatorial test generation tool with support for constraints and higher-index arrays." Paper presented at IEEE International Conference on Software Testing, Verification and Validation ICST 2020 Workshops, Porto, PT. March 23, 2020 - March 27, 2020.

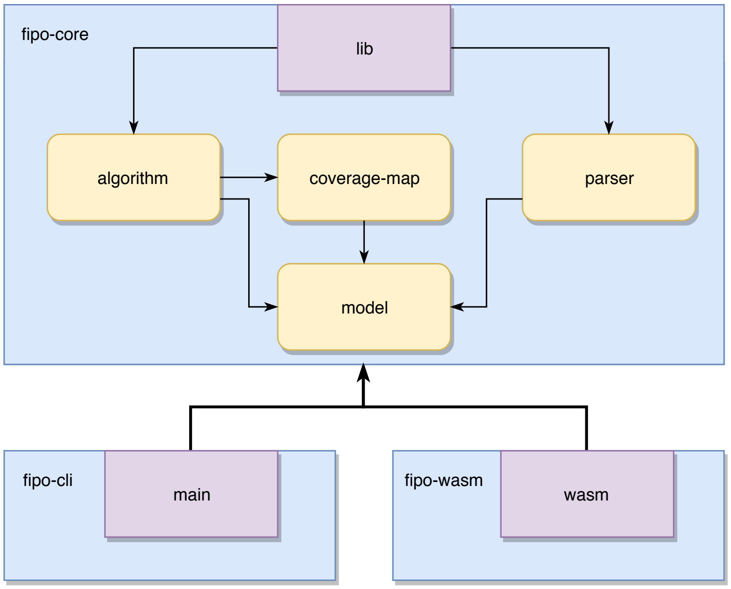


Fig. 1. Rust project structure of CAgen tool

as mathematical and algebraic constructions, exact methods or metaheuristic approaches, have been successfully applied to generate covering arrays [5], but due to the need for flexibility, constraint handling and speed of generation, greedy approaches have proven the most versatile when are applied to real world applications. A popular generation strategy is the one-test-at-a-time approach, where a test set is built by adding one test at a time until all desired t-way interactions are covered. The tests can either be constructed using random methods, as used by AETG [6], or deterministically, such as in the deterministic density algorithm [7] and the algorithm underlying the tool PICT [8]. Another efficient generation strategy is the In-Parameter-Order (IPO) strategy [9]. The strategy starts with a $t \times v^t$ array containing all possible t-way combinations of the first t parameters and continues expanding the array incrementally. In every extension, a horizontal extension is used to add columns (parameters) to the array, while a vertical extension step ensures that the coverage property remains satisfied, adding additional rows (tests) if necessary. While the initially proposed IPO algorithm assigned the values in appended columns greedily from top to bottom, two new variants of the horizontal extension were introduced later on in [10]. The IPOG-F algorithm greedily selects the order in which values are assigned to rows, while the IPOG-F2 algorithm makes use of a heuristic to estimate the gains of selecting a value. All three of these algorithms are available in the ACTS tool as well as in CAgen.

Support for constraints in combinatorial test generation algorithms has been studied intensively [11]. Constraints are a crucial feature for combinatorial test set generation when applied to real-world testing scenarios. The integration of constraints into the test generation process using IPO has been studied previously in [12], [13] and [14]. Currently, two approaches are used in practice: SAT-based and forbidden tuples. The constraint handling techniques used in CAgen are discussed in more detail in Section IV and are based on the forbidden tuples approach, also discussed in [8].

III. Tool Architecture

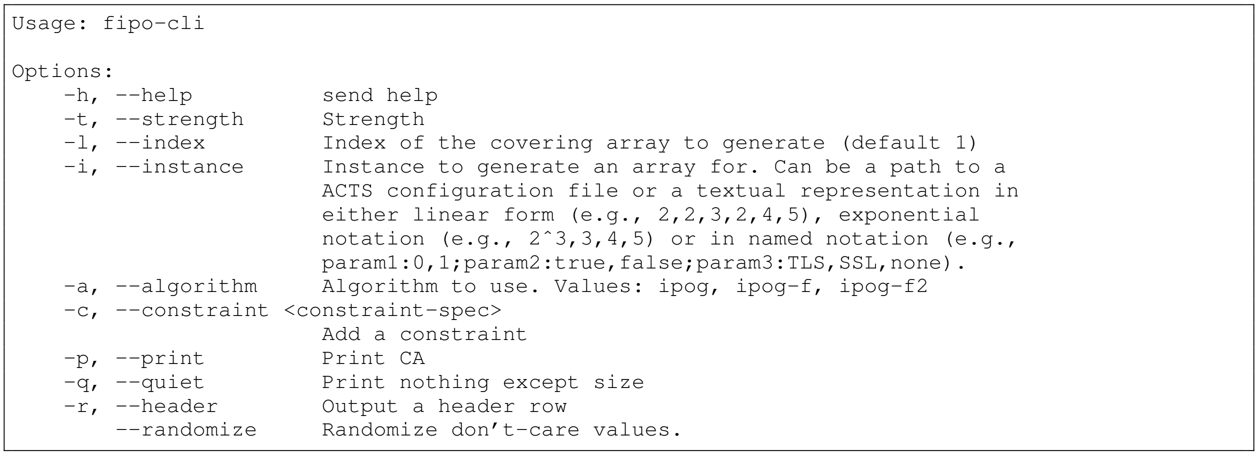
The tool architecture is based on the core algorithm implementations described in [15]. The algorithms are implemented in Rust[2], a fast and safe system-level programming language. The implementation is fast and exhibits a low memory footprint. The basic structure of the project can be found in Figure 1. It is structured into three crates (Rust terminology for package). fipo-core is the library implementing the core algorithms which is further divided into modules. The model module contains common data structures for `Parameter` and `Value` types, the parser module is responsible for parsing common input formats into internal representations. coverage-map and algorithm provide the core data structure and implementations of different IPO algorithms respectively.

Listing 1 shows the interface of the core FIPO algorithm. Two methods for the horizontal and vertical extension can be implemented by different implementations depending on the underlying algorithm. Note that they are parameterized by a generic type T which represents the strength parameter t. Since t is usually a small integer [16], we can instantiate different versions of the algorithm at compile time with constant values for T (e.g., for $t \in \{2, 3, 4, 5, 6, 7, 8\}$) and the compiler can then treat the strength as a constant and perform many useful optimizations such as

[2] https://www.rust-lang.org/en-US/

```
trait CoreIpoAlgorithm {
    fn extend_horizontal<T>(&mut self, column: usize)
        where T: Unsigned;

    fn extend_vertical<T>(&mut self, column: usize)
        where T: Unsigned;

    // non-generic
    fn generate_array(&mut self, strength: usize) {
        // select implementation with given strength
        // -> strength becomes part of the type
        match strength {
            1 => self.generate_with_strength::<U1>(),
            2 => self.generate_with_strength::<U2>(),
            [..]
            _ => panic!(
                format!("t={}␣not␣implemented", strength)
            ),
        };
    }

    fn generate_with_strength<T>(&mut self)
        where T: Unsigned
    {
        // from here on, the strength is part of the type
    }
}
```

Listing 1. Core IPO trait (interface) specification



```
Usage: fipo-cli

Options:
    -h, --help          send help
    -t, --strength      Strength
    -l, --index         Index of the covering array to generate (default 1)
    -i, --instance      Instance to generate an array for. Can be a path to a
                        ACTS configuration file or a textual representation in
                        either linear form (e.g., 2,2,3,2,4,5), exponential
                        notation (e.g., 2^3,3,4,5) or in named notation (e.g.,
                        param1:0,1;param2:true,false;param3:TLS,SSL,none).
    -a, --algorithm     Algorithm to use. Values: ipog, ipog-f, ipog-f2
    -c, --constraint <constraint-spec>
                        Add a constraint
    -p, --print         Print CA
    -q, --quiet         Print nothing except size
    -r, --header        Output a header row
        --randomize     Randomize don't-care values.
```

Fig. 2. Command-line arguments for fipo-cli

]

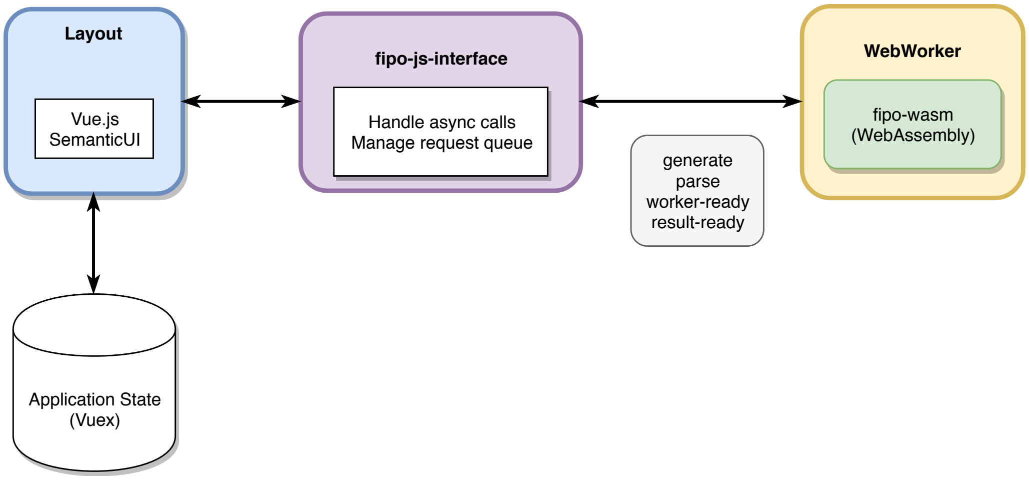


Fig. 3. Architectural overview of CAgen

1) **Compile-time sized t-selections**: t-selections are represented by a constant amount of space in memory and can be tightly packed in an array without necessary meta-data to store the size
2) **Constant iteration bounds**: The loop computing indices for tuples (see definition of *pack* in [15]) can be unrolled, eliminating the inner-most loop of the algorithm altogether

The non-generic `generate_array` method will at run-time select the corresponding generic instantiation of the algorithm based on the strength passed as the argument. As Rust does not yet posses the capability of representing compile-time integers, a type-level encoding using the `typenum`[3] crate is used instead.

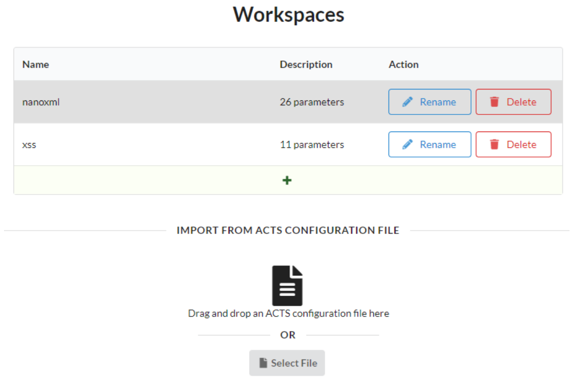
The core of fipo is exposed in two separate frontends.

A. Command-line interface

The command-line interface is implemented using a separate crate (Rust terminology for module) named `fipo-cli`. It is responsible for parsing the given command-line arguments and invoking the desired test generation algorithms for the given configurations. The resulting test sets are then printed to the console according to the formatting wishes of the user. Figure 2 displays all available command line options for `fipo-cli`.

The command-line interface's main use-case is to use it in a broader testing workflow where the tool can be dropped-in to perform (abstract) test case generation.

The main use-case for the command-line interface is to use it as a test generation service which can be embedded into a larger testing work-flow. The interface is kept simple and only allows for the generation of mixed-level covering arrays (MCAs). As such, no explicit test translation (i.e., using model values from an IPM) is provided by `fipo-cli`, but


Workspaces

Name	Description	Action
nanoxml	26 parameters	Rename Delete
xss	11 parameters	Rename Delete

+
IMPORT FROM ACTS CONFIGURATION FILE
Drag and drop an ACTS configuration file here
OR
Select File

Fig. 4. Workspaces manages your IPMs and allows you to import models from config files.

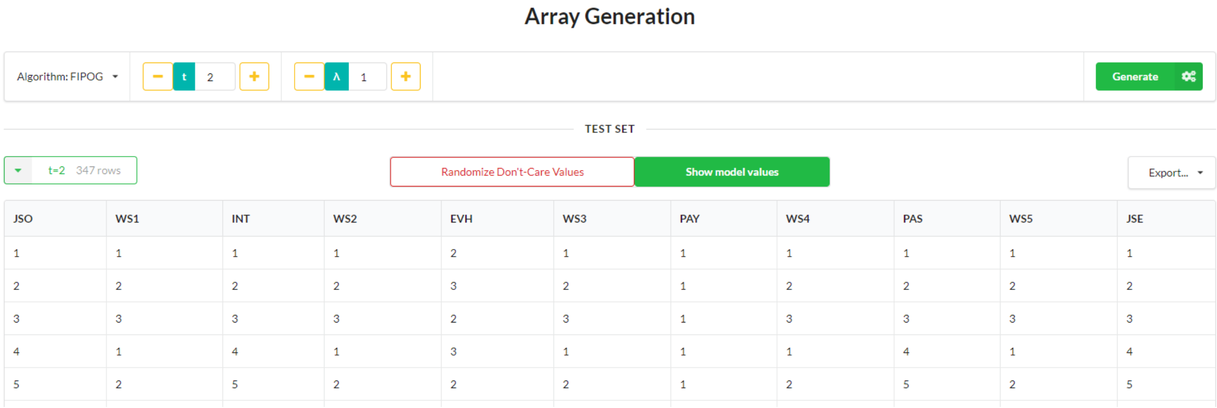
this is trivial to implement by any simple substitution. As test translation often requires additional steps to convert abstract (CA-level) test sets into concrete test sets it makes sense to separate these two parts.

B. Web Application

The web frontend is named CAgen and is a web application bundling the core algorithms of fipo into a user-friendly interface. The general architecture can be found in Figure 3. The most important design decision is the non-requirement of a backend service as the application is completly client-side. It can be hosted on a simple web space and does not require the operator to invest in many computational resources. It is freely available at https://matris.sba-research.org/tools/cagen. The application is a single-page application built with Vue.js[4]

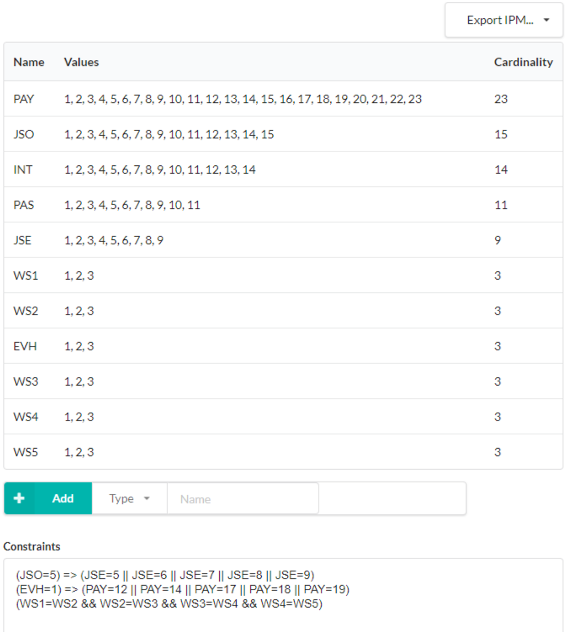
[3] https://crates.io/crates/typenum

[4] https://vuejs.org



JSO	WS1	INT	WS2	EVH	WS3	PAY	WS4	PAS	WS5	JSE
1	1	1	1	2	1	1	1	1	1	1
2	2	2	2	3	2	1	2	2	2	2
3	3	3	3	2	3	1	3	3	3	3
4	1	4	1	3	1	1	1	4	1	4
5	2	5	2	2	2	1	2	5	2	5

Fig. 5. In the generate tab, an algorithm, strength and index can be selected and the resulting array is depicted.



Name	Values	Cardinality
PAY	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23	23
JSO	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15	15
INT	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14	14
PAS	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	11
JSE	1, 2, 3, 4, 5, 6, 7, 8, 9	9
WS1	1, 2, 3	3
WS2	1, 2, 3	3
EVH	1, 2, 3	3
WS3	1, 2, 3	3
WS4	1, 2, 3	3
WS5	1, 2, 3	3

Constraints

```
(JSO=5) => (JSE=5 || JSE=6 || JSE=7 || JSE=8 || JSE=9)
(EVH=1) => (PAY=12 || PAY=14 || PAY=17 || PAY=18 || PAY=19)
(WS1=WS2 && WS2=WS3 && WS3=WS4 && WS4=WS5)
```

Fig. 6. In Input Parameter Model, new parameters can be added easily.

and SemanticUI[5] as the layout framework. The state is managed by Vuex[6]. This state consists mainly of the IPMs managed by the user which are called workspaces in CAgen, see Figure 4. IPMs can be created, edited or uploaded from ACTS files directly in the application and configuration files can even simply be dragged and dropped into the website.

The IPM editor, shown in Figure 6 provides convenience functions for adding boolean, enumeration and range parameters and also supports the addition of parameters given in exponential notation. Further, it is possible to specify and edit constraints for the model.

Lastly, Figure 5 depicts the generate tab, where you can choose between three different algorithms and specify a desired strength and index. Upon clicking the generate button, the tool will generate a covering array for the selected workspace with the specified parameters. Finally, post processing options such as randomizing Don't-Care values as well as an option export the array to different environments are available. For more detailed information and examples, we refer the interested reader to the User Guide in the Help section of our tool.

Providing backend-free test generation: The core of fipo is written in Rust and the language ecosystem has support for the compilation of Rust projects to WebAssembly. WebAssembly is a standardized binary format for executable code supported by all major browsers. It can be used as a common compilation target for different source languages and programs written in these languages can then directly be executed in the browser. Fipo exposes the core algorithms as a separate crate (c.f. Figure 1) which is compiled to WebAssembly and a small JavaScript interface. The web application uses these exposed routines inside a WebWorker which allows for parallel execution (i.e., without blocking the interface with a synchronous call). A small Javascript runner is used to abstract away the worker internals, handle test generation requests asynchronously and also implements queuing.

IV. Main Features

A. *Fast combinatorial test set generation*

A fast run time and low memory usage can play important role in test set generation, as they allow for the generation of test sets for models with a higher number of parameters and values or with higher interaction strength. As an example, CAgen has been one of the key factors for the generation of test sets for testing large-scale data processing at Adobe [1]. In [15], various improvements to the existing IPO algorithms were introduced on the algorithmic as well as on implementation level, briefly summarized as follows. First, it was proposed to compute the coverage gain for same-prefix

[5] https://semantic-ui.com

[6] https://vuex.vuejs.org/en/intro.html

tuples simultaneously, which reduces the reduces the number of operations needed to encode a tuple. Due to the how often this operation is needed, this seemingly small optimization has a significant impact on the run time of the algorithm. Further, storing the number of t-way interactions covered by every column selection in the coverage map makes it possible to exclude fully covered selections from any further considerations. A possible memory optimization is to only keep the currently relevant parts of the coverage map in memory, while the redundant information of the partial CA with $i-1$ columns can be discarded at the begin of the ith horizontal extension step. Another implementation-level optimization is the use of a more efficient vector representation of the array that gets constructed. Lastly, as already mentioned in Section III, the interaction strength can be turned into a compile time constant, enabling various compiler optimizations such as unrolling the loops used to encode tuples.

Section V showcases the speedup these optimizations provide compared to other state-of-the-art tools for the generation of various uniform and mixed level covering arrays.

B. Higher-index CAs

Sometimes it can be desirable that all interactions are covered multiple times instead of at least once, e.g. when multiple activations of an interaction are needed to trigger a fault. With CAgen, it is possible to specify an index λ, where $\lambda = 1$ is equivalent of a traditional covering array. In the generated array, every possible combination is covered at least λ times. We would like to remark, that while this could also be achieved by using the same covering array λ times, our experiments showed that the number of tests needed to cover all interactions at least λ times is usually significantly less than $\lambda \times N$, where N refers to the number of rows in a CA with $\lambda = 1$. For example, for a uniform covering array with 100 columns, $v = 5$ and strength $t = 3$, using IPOG-F, we can construct a $\lambda = 4$ covering array with only 1367 rows, while a normal covering array using the same algorithm has 744 rows.

C. Constraints

CAgen supports user-specified constraints via forbidden tuples. The approach has been described in prior work, however we detail some implementation aspects that have not been covered thus far in the literature in more detail. Specifically we propose a simple way of deriving an initial set of forbidden tuples from arbitrary logical formulas without the need to embed any kind of complex SAT solving mechanism.

The approach is separated into three major steps:

1) Parse user-specified constraints (arbitrary logic formula) into an abstract syntax tree (AST)
2) Visit the AST in post-order to derive an initial set of forbidden tuples
3) Compute minimal forbidden tuples

a) Parsing: The grammar used to parse user-specified constraints can be found in Grammar 1. It allows for arbitrary nesting of formulas using the most commonly used logic operators `or`, `and`, `implication` and `not`. The resulting list of constraints are formulas which must be satisfied by each test in the combinatorial test set. The parser will also check the constraints against the given IPM to make sure that the parameters used in the constraint are part of the model and comparisons with parameter values are with values from the corresponding parameter's domain.

```
⟨constraint-list⟩ ::= ⟨constraint-list⟩ ‘,’ ⟨expr⟩ | ⟨expr⟩

⟨expr⟩ ::= ⟨and-expr⟩
  |  ⟨and-expr⟩ (‘||’ ⟨and-expr⟩)+

⟨and-expr⟩ ::= ⟨impl-expr⟩
  |  ⟨impl-expr⟩ (‘&&’ ⟨impl-expr⟩)+

⟨impl-expr⟩ ::= ⟨unary-expr⟩
  |  ⟨unary-expr⟩ ‘=>’ ⟨unary-expr⟩

⟨unary-expr⟩ ::= ⟨primary-expr⟩ | ‘!’ ⟨unary-expr⟩

⟨primary-expr⟩ ::= ⟨term⟩ | ‘(’ ⟨expr⟩ ‘)’

⟨term⟩ ::= ⟨parameter⟩ ‘=’ ⟨literal⟩
  |  ⟨literal⟩ ‘=’ ⟨parameter⟩
  |  ⟨parameter⟩ ‘!=’ ⟨literal⟩
  |  ⟨literal⟩ ‘!=’ ⟨parameter⟩
  |  ⟨parameter⟩ ‘!=’ ⟨parameter⟩

⟨parameter⟩ ::= r‘[a-zA-Z][a-zA-Z0-9]*’

⟨literal⟩ ::= r‘[0-9]’
  |  ⟨quote⟩ r‘[a-zA-Z0-9_-]+’ ⟨quote⟩
```

Grammar 1. Grammar for user-specified constraints

1) Initial forbidden tuple derivation: Since the constraints are written as positive constraints, i.e. specifying valid configurations, they first need to be converted into a negative representation: forbidden tuples. This step can easily be implemented by visiting each node in the parsed abstract syntax tree in post-order fashion and is defined in Figure 7. Note that p and q represent parameters and v and w corresponding parameter values. $values(p)$ denotes the set of values of the parameter p. Note that in the grammar, the `and` and `or` operators are n-ary instead of binary, but the derive procedure is trivially extendable to handle the $n > 2$ case as well.

Equations (1) - (4) specify base rules to derive forbidden tuples from terms. The resulting set of forbidden tuples for each term is simply the set of all tuples involving the occurring parameters whose value assignments would violate the equation. The remaining rules define the derivation of forbidden tuples for composed expressions. Rules (7) - (11) are simple rewrite-rules for interpreting implications using the `or` and `not` operators as well as pushing negations further down the tree (ending in the negation of a term which is simple to handle according to the first four base rules). This leaves us just with the additional need to handle `or` (5) and `and` (6) operators on the semantic level.

$$\begin{aligned}
derive(p = v) &= \{(p = v') \mid v' \in values(p) \setminus \{v\}\} && (1)\\
derive(p \neq v) &= \{(p = v)\} && (2)\\
derive(p = q) &= \{p = v, q = w) \mid (v, w) \in values(p) \times values(q), v \neq w\} && (3)\\
derive(p \neq q) &= \{p = v, q = w) \mid (v, w) \in values(p) \times values(q), v = w\} && (4)\\
derive(e_1 \,||\, e_2) &= tuple\_product(derive(e_1), derive(e_2)) && (5)\\
derive(e_1 \,\&\&\, e_2) &= derive(e_1) \cup derive(e_2) && (6)\\
derive(e_1 \Rightarrow e_2) &= derive(!e_1 \,||\, e_2) && (7)\\
derive(!!e) &= derive(e) && (8)\\
derive(!(e_1 \,\&\&\, e_2)) &= derive(!e_1 \,||\, !e_2) && (9)\\
derive(!(e_1 \,||\, e_2)) &= derive(!e_1 \,\&\&\, !e_2) && (10)\\
& && (11)
\end{aligned}$$

Fig. 7. Rules for initial tuple derivation

a) and: The formula $e_1 \,\&\&\, e_2$ is not valid if any of the subexpressions e_1 or e_2 are not valid. Thus, any forbidden tuple derived from e_1 and e_2 will be a forbidden tuple for the whole expression and we can just compute the union of the two.

b) or: The formula $e_1 \,||\, e_2$ is not valid if both of the subexpressions evaluate to false. To this end, we must compute the product (`tuple_product`) obtained by joining each tuple from $derive(e_1)$ with all other tuples from $derive(e_2)$ since each tuple derived this way will falsify the entire expression. Note that some tuples can be inconsistent, i.e. contain the same parameter more than once for different value assignments. Such tuples can be discarded since they would not violate the constraint.

2) Minimal Forbidden Tuples Computation: The initial forbidden tuples derived in the previous step are then used to compute the set of minimal forbidden tuples. They are minimal in the sense that removing any parameter value assignment from the tuple will make it valid. This is an important property for test set generation since this means that any test which does not contain a forbidden tuple can be extended without violating any constraint. For more details about the procedure to compute the set of minimal forbidden tuples we refer the reader to [14].

D. Compatibility and Export

CAgen provides a large potential to be used with other combinatorial testing tools by providing different means of exporting both IPMs and generated arrays. IPMs can be exported into formats interpretable by ACTS, CITLAB, PICT and fipo-cli. Arrays can be exported as a comma-separated file (csv) or can directly be exported to Matlab or Python by copying the array into a data structure for the desired target into the users clipboard.

E. Comparison with other GUI tools

Last, we want to compare the feature set provided by the Web-GUI of CAgen with the GUI version of the ACTS tool as well as the service CTWedge. The features of the three test generation tools are summarized in Table I.

a) Features: All three tools support the generation of combinatorial test sets with constraints and provide a way to edit the input parameter models, although CTWedge only provides this in form of a text editor. While the CLI version of ACTS also supports higher strengths, the GUI version is limited to a maximum interaction strength of 6. CAgen supports generation of arrays with strength 8 in both the Web-GUI and the CLI version. As CTWedge is a wrapper around the tools ACTS and CASA, its maximum strength is only limited by the selected tool. Both CAgen and ACTS allow for export of the results and input parameter models into various formats, while CTWedge only supports array exports as .csv files with various seperators. Lastly, only ACTS supports mixed-strength covering arrays and the extension of existing test sets to combinatorial ones, while only CAgen can generate arrays of higher-index and allows you to freely switch between the array with parameter values and the mathematical object underlying it.

b) Distribution: The largest difference between the three tools is the way in which they are distributed. ACTS is, at the time of writing, only available as a pre-compiled Java binary (.jar file) upon request, while both CTWedge and CAgen are accessible via the browser. However, while CTWedge performs the calculations on the server side, CAgen works entirely on the client side using WebAssembly.

c) Privacy: Software testing, by nature, can expose internal details of the system under test and so it might be necessary to keep all testing activity and the produced artifacts in-house. In the case of combinatorial testing, the model might reveal trade-secrets or expose the SUT in other ways so that the chosen testing tool should not share these details with third parties. Both ACTS and CAgen run locally and perform their computations on the users machine. Due to the architecture of CTWedge, the IPM is sent to the server before a test set is returned and thus no such privacy guarantee can be made.

Feature	ACTS	CTWEDGE	CAgen
Max t	6	unlimited	8
IPM editor	GUI	Text Editor	GUI
Constraints	✔	✔	✔
Mixed-strength	✔	✖	✖
Higher-index	✖	✖	✔
Array extension	✔	✖	✖
Show Model	✖	✖	✔
Array export	CSV, NIST, Excel	CSV	CSV, Matlab, Python
IPM export	Text, NIST, XML	??	Text, CITLAB, PICT, fipo-cli
Technology	Java	Java & JS	Rust, WebAssembly & JS
Distribution	Binary	SaaS	Client-side
Privacy	✔	✖	✔

TABLE I
FEATURE SUPPORT

		CAgen		ACTS				PICT			
	t	time(ms)	size	time(ms)	size	time comp.	size comp.	time(ms)	size	time comp.	size comp.
	2	3	347	246	347	**82.00**	**1.00**	15	351	**5.00**	**1.01**
	3	13	4883	373	4896	**28.69**	**1.00**	3741	4898	**287.77**	**1.00**
xss	4	186	54096	2655	54335	**14.27**	**1.00**	633500	54057	**3405.91**	1.00
	5	12215	465011	17872	467754	**1.46**	**1.01**	-	-	-	-
	6	215136	1419206	156192	1436653	0.73	**1.01**	-	-	-	-

TABLE II
RESULTS FOR AN TEST SET WITH CONSTRAINTS.

V. PERFORMANCE BENCHMARKS

To evaluate the performance of the IPOG algorithm implementation of CAgen, we compare its CLI version against the implementation of the same algorithm in the CLI version of ACTS, as well as with Microsoft's pairwise testing tool PICT. For each considered instance, we report the run time as well as the size of the generated arrays. Note that, while ACTS and CAgen use the same algorithm, the size of the resulting arrays varies slightly due the tie breaker selection. For an in-depth analysis of tiebreakers for algorithms of the IPO family, we refer the interested reader to [17]. For CAgen and PICT, the Linux time command was used for measuring the run times, while for ACTS we used the run time reported by the tool, as the overhead produced by java would not allow an accurate algorithmic comparison for smaller instances. All experiments were conducted on a server equipped with 64 GB of RAM and an Intel Xeon E3 processor.

First, we tested the generation of uniform CAs with 10 columns, strengths $t = 2$ to $t = 4$ and values arising from an alphabet with cardinality between 10 and 100. The results of our experiments are depicted in Table III. For strength two, CAgen seems to achieve a consistent speedup between 20 and 30 compared to ACTS, while producing slightly smaller arrays on average. For higher strengths, we can notice the run time difference increase with higher alphabets, with CAgen generating the desired CAs more than 100 times faster for high alphabets. The algorithm underlying PICT seems to scale even worse with bigger instance sizes, which made it impossible to generate arrays for many of the tested instances within reasonable time.

Afterwards we generated test sets for models from real world applications, also used in [18] and [19] , shown in Table IV. Once again, all instances report a significant speed up compared to ACTS, while generally generating slightly smaller arrays. PICT seems to be competitive in terms of speed for smaller $t = 2$ instances, but once the strength is increased, it becomes too slow for practical use.

Lastly, we tested a model for exploiting web application security vulnerabilities (XSS) that contains various constraints. The model and it's application is explained in detail in [20] and the results are depicted in Table II. For smaller strengths, CAgen still finishes significantly faster than ACTS, while for high strengths, ACTS manages to generate CAs slightly faster, suggesting that our constraint implementation still leaves room for improvement.

VI. CONCLUSION AND FUTURE WORK

In this work we presented CAgen, a tool for generating combinatorial test sets. It comes as a Web-GUI or a command line interface and contains various features necessary for combinatorial testing, such as support for higher strengths, constraints and covering arrays of higher-index. Our benchmarks show that CAgen can construct covering arrays significantly faster than other state-of-the-art tools such as ACTS or PICT, without compromise in the number of test cases needed. In the future, we want to increase the number of features CAgen supports even further, by enabling an option to extend existing test sets to combinatorial ones. Furthermore, in [21], better support for negative testing was suggested, which can be done to some degree manually, using constraints, but is still something we want to include in future releases of the tool. Lastly, in addition to the currently provided efficient greedy algorithms, we want to extend CAgen by adding support for various metaheuristic approaches [22] and hybrid heuristics [23] as well as algorithms based on learning methods [24]. Thus, we envision that in the future CAgen will become an all-around framework useful to developers as well as researchers.

		CAgen		ACTS				PICT			
t	v	time(ms)	size	time(ms)	size	time comp.	size comp.	time(ms)	size	time comp.	size comp.
2	10	1	160	20	157	**20.00**	0.98	9	166	**9.00**	**1.04**
	20	3	599	86	604	**28.67**	**1.01**	97	662	**32.33**	**1.11**
	30	5	1299	125	1308	**25.00**	**1.01**	446	1336	**89.20**	**1.03**
	40	9	2277	224	2265	**24.89**	0.99	1334	2310	**148.22**	**1.01**
	50	15	3522	401	3522	**26.73**	**1.00**	3164	3539	**210.93**	**1.00**
	60	23	4978	604	4969	**26.26**	1.00	6433	5025	**279.70**	**1.01**
	70	34	7332	975	7325	**28.68**	1.00	11632	6759	**342.12**	0.92
	80	44	8609	1185	8615	**26.93**	**1.00**	19641	8751	**446.39**	**1.02**
	90	62	10878	1760	10891	**28.39**	**1.00**	31047	10962	**500.76**	**1.01**
	100	88	13307	1954	13329	**22.20**	**1.00**	47749	13463	**542.60**	**1.01**
3	10	38	2360	158	2377	**4.16**	**1.01**	2818	2324	**74.16**	0.98
	20	123	18101	1852	18167	**15.06**	**1.00**	163211	17354	**1326.92**	0.96
	30	584	59299	8496	59589	**14.55**	**1.00**	1789929	56386	**3064.95**	0.95
	40	1468	137585	42637	138390	**29.04**	**1.01**	10178576	130173	**6933.63**	0.95
	50	3514	265408	136378	266484	**38.81**	**1.00**	-	-	-	-
	60	7094	451356	363621	454877	**51.26**	**1.01**	-	-	-	-
	70	13939	710589	917330	715400	**65.81**	**1.01**	-	-	-	-
	80	22715	1048958	2261417	1057705	**99.56**	**1.01**	-	-	-	-
	90	38367	1486210	4083039	1495654	**106.42**	**1.01**	-	-	-	-
	100	56318	2017231	6672042	2038072	**118.47**	**1.01**	-	-	-	-
4	10	260	29612	5195	29466	**19.98**	1.00	761686	28036	**2929.56**	0.95
	20	9500	458607	389087	454139	**40.96**	0.99	-	-	-	-
	30	125364	2271737	8739068	2245317	**69.71**	0.99	-	-	-	-
	40	1335707	7070939	-	-	-	-	-	-	-	-
	50	8902513	17059586	-	-	-	-	-	-	-	-

TABLE III
RESULTS OF GENERATED UNIFORM CAS.

		CAgen		ACTS				PICT			
	t	time(ms)	size	time(ms)	size	time comp.	size comp.	time(ms)	size	time comp.	size comp.
mobile	2	19	157	78	155	**4.11**	0.99	133	180	**7.00**	**1.15**
	3	505	2236	2805	2243	**5.55**	**1.00**	186647	2240	**369.60**	**1.00**
	4	74011	27538	1349939	27561	**18.24**	**1.00**	-	-	-	-
	5	17303556	304887	-	-	-	-	-	-	-	-
wireless	2	5	45	19	47	**3.80**	**1.04**	4	47	0.80	**1.04**
	3	30	315	64	302	**2.13**	0.96	441	306	**14.70**	0.97
	4	147	1841	951	1839	**6.47**	1.00	58191	1824	**395.86**	0.99
	5	3148	11064	38960	10169	**12.38**	0.92	5801222	10002	**1842.83**	0.90
	6	62125	57633	801882	52515	**12.91**	0.91	-	-	-	-
flex	2	3	26	18	26	**6.00**	**1.00**	3	26	**1.00**	**1.00**
	3	14	91	52	97	**3.71**	**1.07**	83	99	**5.93**	**1.09**
	4	45	347	234	361	**5.20**	**1.04**	4696	362	**104.36**	**1.04**
	5	572	1164	4216	1171	**7.37**	**1.01**	198509	1158	**347.04**	0.99
	6	7201	3527	100539	3547	**13.96**	**1.01**	-	-	-	-
make	2	1	30	17	31	**17.00**	**1.03**	2	32	**2.00**	**1.07**
	3	3	138	46	141	**15.33**	**1.02**	67	149	**22.33**	**1.08**
	4	36	607	157	608	**4.36**	**1.00**	3551	619	**98.64**	**1.02**
	5	279	2208	1351	2241	**4.84**	**1.01**	130499	2215	**467.74**	**1.00**
	6	2473	7153	28159	7289	**11.39**	**1.02**	-	-	-	-
grep	2	5	273	24	275	**4.80**	**1.01**	13	273	**2.60**	**1.00**
	3	9	2732	95	2739	**10.56**	**1.00**	1242	2732	**138.00**	**1.00**
	4	92	19175	408	19406	**4.43**	**1.01**	122053	19215	**1326.66**	**1.00**
	5	620	97024	3635	97879	**5.86**	**1.01**	6053545	99471	**9763.78**	**1.03**
	6	5113	406859	91057	418645	**17.81**	**1.03**	-	-	-	-
sed	2	5	81	22	84	**4.40**	**1.04**	9	86	**1.80**	**1.06**
	3	20	679	82	679	**4.10**	**1.00**	461	667	**23.05**	0.98
	4	103	4546	413	4608	**4.01**	**1.01**	55253	4561	**536.44**	**1.00**
	5	1169	25846	7107	27697	**6.08**	**1.07**	4613703	27530	**3946.71**	**1.07**
	6	17323	143638	218384	149289	**12.61**	**1.04**	-	-	-	-
gzip	2	4	204	25	206	**6.25**	**1.01**	16	204	**4.00**	**1.00**
	3	12	1038	80	1094	**6.67**	**1.05**	1266	1081	**105.50**	**1.04**
	4	172	5138	809	5284	**4.70**	**1.03**	121923	5367	**708.85**	**1.04**
	5	2933	23690	10605	23871	**3.62**	**1.01**	8193199	23625	**2793.45**	1.00
	6	42749	93674	292711	95502	**6.85**	**1.02**	-	-	-	-
nanoxml	2	4	25	16	26	**4.00**	**1.04**	7	26	**1.75**	**1.04**
	3	8	102	30	105	**3.75**	**1.03**	41	110	**5.13**	**1.08**
	4	44	387	100	398	**2.27**	**1.03**	1440	398	**32.73**	**1.03**
	5	149	1320	594	1287	**3.99**	0.98	41439	1304	**278.11**	0.99
	6	956	4183	6428	3973	**6.72**	0.95	999012	3935	**1044.99**	0.94

TABLE IV
RESULTS OF GENERATED MCAS FOR TESTING REAL WORLD APPLICATIONS.

Acknowledgement

The research presented in this paper has been funded in part by the Austrian Research Promotion Agency (FFG) under grant 865248 (SECuring Web technologies with combinatorial Interaction Testing - SecWIT) and in the context of the Austrian COMET K1 program.

Moreover, this work was performed partly under the following financial assistance award 70NANB18H207 from U.S. Department of Commerce, National Institute of Standards and Technology.

Disclaimer: *Products may be identified in this document, but identification does not imply recommendation or endorsement by NIST, nor that the products identified are necessarily the best available for the purpose.*

References

[1] R. Smith, D. Jarman, R. Kacker, R. Kuhn, D. Simos, L. Kampel, M. Leithner, and G. Gosney, "Applying combinatorial testing to large-scale data processing at adobe," in *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, April 2019, pp. 190–193.

[2] A. Gargantini and M. Radavelli, "Migrating combinatorial interaction test modeling and generation to the web," in *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2018, pp. 308–317.

[3] J. Bonn, K. Foegen, and H. Lichter, "A framework for automated combinatorial test generation, execution, and fault characterization," in *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, April 2019, pp. 224–233.

[4] L. Yu, Y. Lei, R. N. Kacker, and D. R. Kuhn, "Acts: A combinatorial test generation tool," in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation*. IEEE, 2013, pp. 370–375.

[5] J. Torres-Jimenez, I. Izquierdo-Marquez, and H. Avila-George, "Methods to construct uniform covering arrays," *IEEE Access*, vol. 7, pp. 42 774–42 797, 2019.

[6] D. M. Cohen, S. R. Dalal, M. L. Fredman, and G. C. Patton, "The aetg system: An approach to testing based on combinatorial design," *IEEE Transactions on Software Engineering*, vol. 23, no. 7, pp. 437–444, 1997.

[7] R. C. Bryce and C. J. Colbourn, "The density algorithm for pairwise interaction testing," *Software Testing Verification and Reliability*, vol. 17, no. 3, pp. 159–182, 2007.

[8] J. Czerwonka, "Pairwise testing in real world," in *24th Pacific Northwest Software Quality Conference*, vol. 200. Citeseer, 2006.

[9] Y. Lei and K.-C. Tai, "In-parameter-order: A test generation strategy for pairwise testing," in *High-Assurance Systems Engineering Symposium, 1998. Proceedings. Third IEEE International*. IEEE, 1998, pp. 254–261.

[10] M. Forbes, J. Lawrence, Y. Lei, R. N. Kacker, and D. R. Kuhn, "Refining the in-parameter-order strategy for constructing covering arrays," *Journal of Research of the National Institute of Standards and Technology*, vol. 113, no. 5, pp. 287–297, 2008.

[11] H. Wu, C. Nie, J. Petke, Y. Jia, and M. Harman, "A survey of constrained combinatorial testing," *ArXiv*, vol. abs/1908.02480, 2019.

[12] L. Yu, Y. Lei, M. Nourozborazjany, R. N. Kacker, and D. R. Kuhn, "An efficient algorithm for constraint handling in combinatorial test generation," in *2013 IEEE Sixth International Conference on Software Testing, Verification and Validation*, 2013, pp. 242–251.

[13] L. Yu, F. Duan, Y. Lei, R. N. Kacker, and D. R. Kuhn, "Combinatorial test generation for software product lines using minimum invalid tuples," in *High-Assurance Systems Engineering (HASE), 2014 IEEE 15th International Symposium on*. IEEE, 2014, pp. 65–72.

[14] ——, "Constraint handling in combinatorial test generation using forbidden tuples," in *2015 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 2015, pp. 1–9.

[15] K. Kleine and D. E. Simos, "An efficient design and implementation of the in-parameter-order algorithm," *Mathematics in Computer Science*, vol. 12, no. 1, pp. 51–67, 2018.

[16] D. R. Kuhn, R. N. Kacker, and Y. Lei, *Introduction to Combinatorial Testing*, 1st ed. Chapman & Hall/CRC, 2013.

[17] K. Kleine, I. Kotsireas, and D. E. Simos, "Evaluation of tie-breaking and parameter ordering for the ipo family of algorithms used in covering array generation," in *Combinatorial Algorithms*, C. Iliopoulos, H. W. Leong, and W.-K. Sung, Eds. Springer International Publishing, 2018.

[18] J. Petke, M. B. Cohen, M. Harman, and S. Yoo, "Practical combinatorial interaction testing: Empirical findings on efficiency and early fault detection," *IEEE Transactions on Software Engineering*, vol. 41, no. 9, pp. 901–924, Sep. 2015.

[19] S. A. Seidel, K. Sarkar, C. J. Colbourn, and V. R. Syrotiuk, "Separating interaction effects using locating and detecting arrays," in *Combinatorial Algorithms*, C. Iliopoulos, H. W. Leong, and W.-K. Sung, Eds. Cham: Springer International Publishing, 2018, pp. 349–360.

[20] J. Bozic, B. Garn, D. E. Simos, and F. Wotawa, "Evaluation of the ipo-family algorithms for test case generation in web security testing," in *2015 IEEE Eighth International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, April 2015, pp. 1–10.

[21] C. Eitner and F. Wotawa, "Crucial tool features for successful combinatorial input parameter testing in an industrial application," in *2019 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, April 2019, pp. 188–189.

[22] M. Wagner, L. Kampel, and D. E. Simos, "Quantum-inspired evolutionary algorithms for covering arrays of arbitrary strength," in *Analysis of Experimental Algorithms*, I. Kotsireas, P. Pardalos, K. E. Parsopoulos, D. Souravlias, and A. Tsokas, Eds. Cham: Springer International Publishing, 2019, pp. 300–316.

[23] ——, "IPO-Q: A quantum-inspired approach to the ipo strategy used in ca generation," in *To appear in Proceedings of Mathematical Aspects of Computer and Information Sciences 2019*. Springer International Publishing, 2019.

[24] L. Kampel, M. Wagner, I. S. Kotsireas, and D. E. Simos, "How to use boltzmann machines and neural networks for covering array generation," in *To appear in Proceedings of the 13th Lion Leraning & Intelligent Optimization Conference*, May 2019, pp. 32–46.

Proceedings of the ASME 2020 International Manufacturing Science and Engineering Conference, MSEC 2020

June 22-26, 2020, Cincinnati, OH, USA

MSEC2020-12885

MSEC: A QUANTITATIVE RETROSPECTIVE

Thurston Sexton
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Michael P Brundage*
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Alden Dima
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

Michael Sharp
National Institute of Standards and Technology
Gaithersburg, MD 20814, USA

ABSTRACT

The ASME 2020 Manufacturing Science and Engineering Conference (MSEC) is the 15th annual meeting organized by the Manufacturing Engineering Division (MED) of ASME. MED and ASME MSEC focuses on manufacturing sciences, technology, and applications, including machining, materials processing, sensing, robotics, manufacturing system dynamics, and production optimization. As the conference has grown and evolved from its inception, it can be difficult to intuitively visualize and discuss the broad range of research topics covered by the MSEC community or to intuitively ascertain their evolution throughout time. This paper discusses a methodology to quantitatively model research communities within bodies of literature—specifically, the relative change of relevant topics within AMSE MSEC conference papers through time, from 2006 through 2018. The goal of this work is to not only present how research in MSEC has shifted over time, but in a broader sense to provide a discussion on how others can interpret results so that similar analysis can be produced within other research communities. This methodology can be used to identify overlap of communities, monitor growth or stagnation within the communities, to aid in developing new symposiums and communities of interest, or even to dictate future standards needs by looking at research trends and subsequent standard development.

1 INTRODUCTION

In 2020, the Manufacturing Engineering Division (MED) in ASME (The American Society of Mechanical Engineers) will observe its 100 year anniversary [1]. MED focuses on “the knowledge base of manufacturing sciences and technology and its applications for improved production performance that is economically viable and meets industrial health, safety, and resource conservation legislation” [2]. In conjunction with its 100 year anniversary, MED will hold the 15th annual Manufacturing Science and Engineering Conference (MSEC) in Cinncinati, OH. Traditionally, MED and MSEC have covered a wide array of manufacturing topics, including: machine tools, materials processing, sensors and controllers, computer integrated manufacturing and robotics, manufacturing systems management and optimization, and emerging areas of manufacturing engineering [2].

As published research in manufacturing evolves, a method is needed to quantitatively evaluate the research topics within communities and interpret their evolution over time. This paper uses established statistical modeling and natural language processing (NLP) techniques, such as topic modeling and document cluster tracking, with papers from ASME MSEC on 1) the topics of discussion within ASME MSEC, and 2) how related research efforts have evolved over time within those topics. The goal of this paper is to not only discuss interesting trends in ASME MSEC, but more importantly to provide a detailed process for obtaining and interpreting these results. The larger goal is to ensure that the developed procedure can be replicated for other research collec-

*Corresponding Author: mpb1@nist.gov

Sexton, Thurston; Brundage, Michael; Dima, Alden A.; Sharp, Michael. ”MSEC: A Quantitative Retrospective.” Paper presented at MSEC: Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

tions, providing needed insights for other communities and domains. This quantitative method for evaluating and interpreting the evolution of research communities and their primary topics of interest can be used to identify success stories in areas of rapid growth or movement, new and developing thrusts that may be worthy of additional attention, stagnant topics in need of revitalization, as well as predicting future needs and next steps. Information gained through these types of analysis within a specific domain could even be used to help describe future conference tracks or symposiums, highlight standards needs, and provide justification for potential future research funding areas.

The rest of the paper is structured as follows. Section 2 discusses background on topic modeling, document cluster tracking, and other literature discussing research paper trends. Section 3 illustrates the methodology used to create the topics and document shift results. Lastly, conclusions and future work are presented in Section 4.

2 PROBLEM CONTEXT

With the rapidly expanding volume of available documents in a given research area, it is nearly impossible for a single person to manually survey and synthesize a full community of research. Evolution of language usage as well as changes in interest and focus through time can make identifying and evaluating groups of common research difficult. This paper presents a method to quantitatively assess and follow groups of semantically similar documents to create a timeline of research thrusts that identifies progressive changes in interest, and the relative interplay between thrust areas. Additionally, results are presented to gain a more global view of the intrinsic topic areas that arise through time and how different ideas will come in and out of vogue across the corpus of documents.

2.1 Data Acquisition

The quality of any analysis is directly related to the quality of the data used in the analysis, in this case the presented data relates a sampling of the available MSEC documents from 2006 to 2018. Information contained in the ASME Digital Collection web pages was used to identify and download MSEC conference papers as individual PDF files using the workflow created by the authors, shown in Fig.1. This step produced the data set used in our analysis by extracting and processing their text from the pdfs. However, due to a lack of a standardized storage and access format, the number of papers available for this research between 2006 and 2012 is significantly lower that the full number of papers identified by citation, as shown in Fig. 2. The work presented in Section 3 still includes these years for completeness and for demonstration of the method, but any conclusions drawn from this region must be regarded as highly uncertain due to the low number of full text papers that could be obtained for this work.

The top-level conference-related pages list conference papers by track and symposium. Starting from a list of manually collected top-level URLs, the data gathering workflow used a combination of common Unix command-line tools and established Python libraries to download the top-level conference papers (Fig. 1: Web Page Downloader). The downloaded pages were processed to extract the article metadata directly from the page HTML (Fig. 1: Metadata Extraction) including: the conference name, the PDF URL, and the article ID, DOI, Date, and URL. Each article's track and symposium was then extracted from the structural information of the conference proceedings webpage (Fig. 1: Track & Symposium Extraction). Using the extracted data, we identified a total of 2267 articles and were able to download 1457 MSEC-related PDFs from the ASME Digital Collection (see Fig. 2 and Fig. 1: PDF Downloader). We extracted their text, including titles, table and figure captions, and metadata, via the pdftotext command-line tool (Fig. 1: Text Extraction).

Though there are software tools that can be used to automate portions of the data acquisition, the general form of the task of obtaining and collating a collection of documents will still require human supervision and intervention at each stage. Various factors like the form and location of the publications, possibly Web site structure, document layout, and file naming conventions all affect the ability to automate the discovery and extraction of articles. There is also strong possibilities of desirable tacit information being encoded that is incredibly difficult for an automated algorithm to address without special considerations being made.

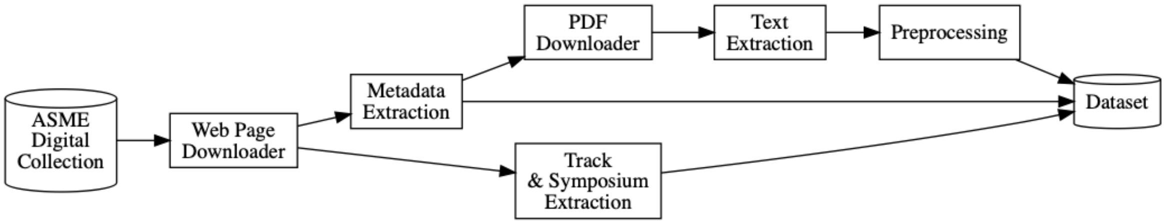


FIGURE 1. *Workflow used to acquire and process ASME MSEC articles.*

During our study, important information about the relationship between articles and symposia were encoded structurally in the Web pages and required human inspection and analysis to write the custom code for its extraction. Though each instance of acquiring data to process will have unique eccentricities, the most important aspects of collection for our pipeline are to ensure that each document with computer interpretable text is accessible, and their associated publication date are linked to them. Any additional meta data about locations, tracts, authorship, etc. can be useful for additional analyses, or may facilitate the data acquisition, but are not critically needed for the described analysis.

2.2 Data Preparation

Even in digital form, text is not amenable to direct analysis, especially by mathematical techniques based on linear algebra, which are typical of many NLP algorithms. Preprocessing is necessary to convert the text into a more suitable representation. We will discuss text preprocessing in the following two sections. Section 2.2.1 will give an overview of the different facets of text preprocessing. Section 2.2.2 will describe how we prepared our text for the analyses described in Section 3.

2.2.1 Overview of Text Preprocessing

Text is often preprocessed before analysis, with one goal being to simplify the data for analysis without losing necessary information [3]. This simplification can be thought of as sharpening the desired "signal" while reducing "noise". Another simultaneous goal is to structure the data in a manner that facilitates the analysis [4]. This restructuring is often designed to emphasize certain features that are important to the analysis and algorithms at the expense of others that may be less important. Examples of typical preprocessing beneficial to these types of analysis include:

Extraneous Metadata Removal: The removal of unwanted template text [5] that add "noise" to analyses. For example, each page contained text such as:

> Downloaded from https://asmedigitalcollection.asme.org
> /MSEC/proceedings-pdf/MSEC2008/48517/1/2715771
> /1_1.pdf by NIST user on 18 September 2019

This unwanted text can be identified for removal using technologies such as regular expressions, a mathematical language for describing text patterns used by text processing software to identify specific text [6].

Stop Word Removal: The elimination of many common words such as *this, that,* and *the* that convey little semantic meaning to an analysis [7]. Preassembled *Stop Lists* can be used to identify and remove them from text prior to analysis [7, 8].

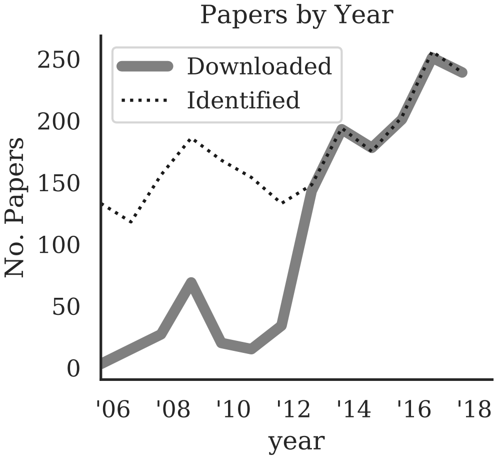


FIGURE 2. *Papers identified and downloaded from the ASME Digital Collection by year using our workflow*

Lemmatization: The reduction of inflected words via linguistic techniques into their dictionary forms [9], for example:

$$\{\text{boat, boats, boat's, boats'}\} \Rightarrow \text{boat} \tag{1}$$

By mapping all of the variants of a word into its base form, lemmatization can be seen as a way to increase the "signal" associated with that word.

Cleaning: The removal of characters and tokens, such as punctuation, numbers, and email addresses as well as low-frequency words, from the text to reduce noise features in the data that can negatively impact analysis [3, 9]

Tokenization: The process of separating text into its smallest meaningful units such as words and numbers [4, 7]. Because it identifies individual words, tokenization is a fundamental process that serves other downstream preprocessing and analysis tasks.

Sentence Segmentation: The identification of sentence boundaries in a text [4, 8]. Sentence segmentation plays a key role in supporting lemmatization by providing the input for the linguistic analyses required for lemmatization.

Bag of Words (BOW): A method for representing the contents of documents by identifying their words and counts without any ordering information [9]. For example:

$$\text{“Jack has a blue hat and Jill has a red hat”} \Rightarrow \{(\text{“Jack”},1),(\text{“Jill”},1),(\text{“a”},2),(\text{“and”},1), (\text{“blue”},1),(\text{“has”},2),(\text{“hat”},2),(\text{“red”},1)\} \quad (2)$$

Vectorization: In the presence of a dictionary to assign unique integers to each word, BOW can also be represented as vectors. For example, assuming that the previously used sentence was the first sentence in a text collection then the dictionary would contain the following information:

$$\text{dict} = \{(0,\text{“Jack”}),(1,\text{“has”}),(2,\text{“a”}),(3,\text{“blue”}), (4,\text{“hat”}),(5,\text{“and”}),(6,\text{“Jill”}),(7,\text{“red”})\} \quad (3)$$

The integers serve as word ids and as subscripts to the vector representations for the BOW. The first example now becomes:

$$\text{“Jack has a blue hat and Jill has a red hat”} \Rightarrow (1,2,2,1,2,1,1,1,0,\ldots) \quad (4)$$

where $0,\ldots$ represents positions associated with other words in the text collection not found in the first sentence. In practice, these vectors are sparse; for a typical text collection, zeros account for about 98% of the vector elements [10].

Considerations Text preprocessing is a form of analysis on its own, where typically the actual analysis is hidden from the user and performed in preexisting libraries and frameworks. For example, lemmatization requires the identification of the parts of speech that also depend on the identification of individual sentences. These operations are implemented by the NLP framework (Section 2.2.2) and involve commonly available statistical and neural-network language models whose full descriptions are outside the scope of this paper. However, interested readers can refer to books such as Krohn et al [11] and Rao and McMahon [12] and articles such as those by Young et al [13] and Akbik et al [14]. The ultimate text analysis performed thus rests upon the these hidden analyses.

While it is convenient to visualize text preprocessing as a pipeline in which raw text flows in and is replaced with refined text for analysis, it is better to initially consider it as a series of stages in an converging iterative process. Ideally, the entire output of each stage should be available for inspection and analysis; “sanity checks” are crucial. The pipeline perspective makes the most sense at the end when all of the issues discovered during development have been resolved.

2.2.2 Text Preprocessing Used

Our final data set contained the following preprocessing steps. First, filenames were standardized to contain both the year and a unique paper ID for easy of storage, indexing, and retrieval. We then filtered the extracted text from each document using Python’s regular expression library to remove headers and extraneous metadata that can negatively impact the subsequent text analysis. For example, each PDF file contained text identifying its source URL, the account that downloaded it, and the date and time that it was downloaded; none of which is relavant to the subject of the document. The text was then segmented into individual sentences and tokenized words. *Stop words* were discarded and the remaining tokens were lemmatized using established procedures found in `spaCy`[1]. Special word/number occurrences were replaced with unique text-only aliases prior to the removal numeric entities. For example:“1D”, “2D”, and “3D” were replaced with “oned”, “twod”, and “threed”, respectively. The text was then scrubbed using the `clean-text` library[2] to reduce it to lower case and to remove URLs, email addresses, phone numbers, currency symbols, punctuation, and numbers. At the end of this phase of data preparation, each downloaded PDF document had a corresponding text file that contained a single processed sentence per line.

We then used gensim[3] to build a look up table style dictionary that relates individual tokens to a unique numerical identifier and converted each text file into a BOW represented by gensim’s sparse vector format [15]. These processes have allowed the data to now be ingested and analyzed by modern mathematical analysis techniques.

The presence of semantically interesting non-word tokens (such as “3D” and its replacement with “threed”) highlights that text preprocessing is transformative in nature. While it may be intuitive that some text should be removed as being not relevant, it may be less obvious that text analysis relies on the translation of the original text, for example the use of lemmatization, substitutions, and bags of words. Text preprocessing should not be seen as the rote application of recipes to the original text. Instead analysts must inspect the intermediate results while considering their ultimate goals. For a discussion of how preprocessing can alter analysis, see Denny and Spirling [3].

3 ANALYSIS WORKFLOW & RESULTS

To streamline analysis and interpretation of collections of research documents, this paper presents a workflow for deriving useful insights from raw text, not only from the collections gathered at MSEC, but in other domains as well. Many options exist for the mechanical implementation of each of the steps in the process, but the choice of specific algorithms or combinations of algorithms has the potential to greatly influence the quality and

[1] `https://spacy.io/`
[2] `https://github.com/jfilter/clean-text`
[3] https://radimrehurek.com/gensim/

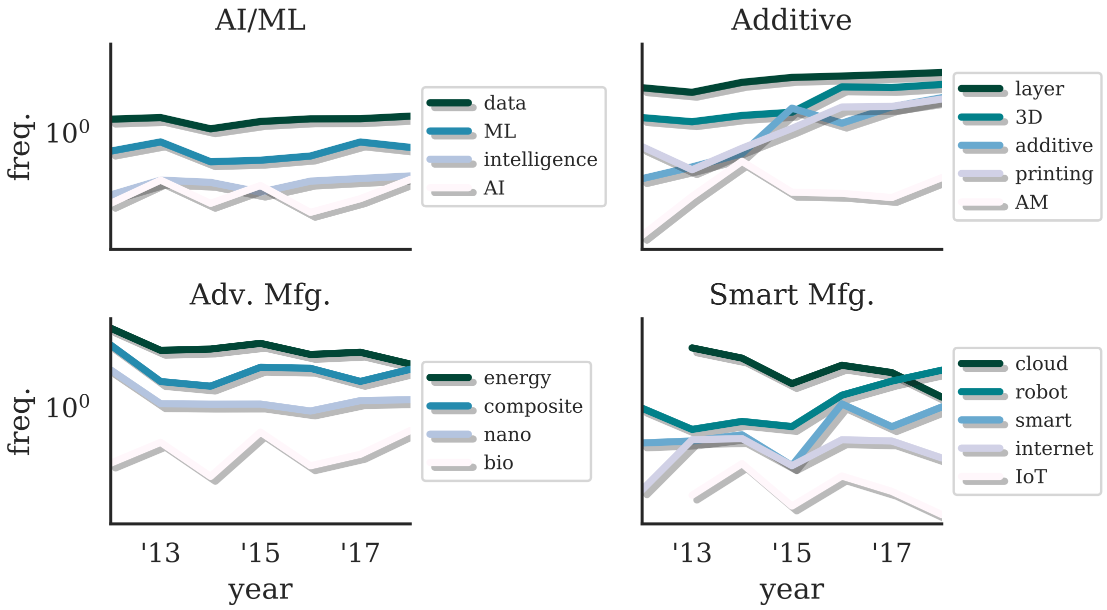


FIGURE 3. Keyword occurrences per document (log-scale) over time, for documents submitted after 2012. Each figure shows the top most-frequent terms occurring within the author-determined research areas. Darker color represents higher all-time occurrence.

outcome of the end analysis. This section discusses the benefits and difficulties of choices made at each step of the process to develop a guideline and philosophical approach for readers wishing to apply the procedure to better understand their own domain.

3.1 Keyword Trends

One *direct* way to access trends in a domain over time is to observe how keyword usage evolves within the documents that make up the domain. Given a frame of reference, such as a defined area of interest (*e.g.* , guiding documents from steering committees) we can categorize keywords by their relevant topic and observe the relative frequency of these keywords within.

For example, ASME defined several key research areas[4] that may define the coming decade in engineering research. Several of these are highly relevant to manufacturing specifically. By selecting keywords typically associated with each area, whether from anecdotal experience or expertise, we can observe high-level trends in the focus each area has over time.

Figure 3 illustrates an example selection of these key research areas identified from ASME with several terms the authors deem as relevant to those areas. Each term is reported for the years in which a significant fraction of total papers were recovered to avoid bias. Importance is compared through total expected occurrences *per paper*.

Note that this analysis only shows general trends over time– whether a word is used more or less frequently at a global scale. In *AI/ML* we see a slight increase in “data” over time with “Machine Learning” displaying an earlier surge than “Artificial Intelligence”, which surged recently after a slight down-turn in use in 2015-2016. *Additive* sees a steady increase across all major keywords, likely due to an overall increase of total interest (paper submissions) in the topic. Although a fairly broad topic scope, *Advanced Manufacturing* is seeing continued research in nano-scale and composites fabrication. A long-term decline in “energy” interest is being met with a rise in “bio-manufacturing” interest. Finally, *Smart Manufacturing* is showing a marked increase in “robot” use, while words like “Internet of Things” and “cloud” seem to be declining from a peak near 2014.

While this may present an excellent first-look into the trends within a community-of-interest, the actual “dynamics” here can be difficult to parse. Despite the apparent correlation between certain keywords, the same could be said about terms across topics. This is problematic if one wishes to make quantitative generalizations about topics overall, or even how well individual terms reflect the state of a topic. Do these “areas of interest” constitute an efficient snapshot of the community as a whole, and the trends hidden within?

[4] https://www.asme.org/wwwasmeorg/media/resourcefiles/aboutasme/asme_strategy.pdf

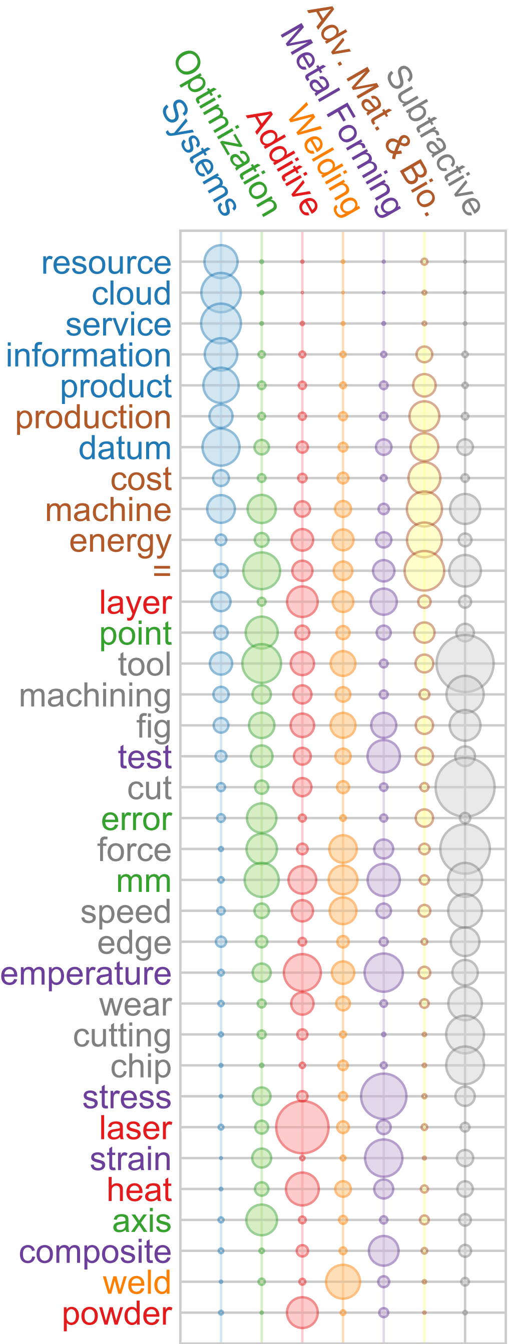


FIGURE 4. Term probabilities within each of 7 topics, as named by the authors. Showing top 35 terms having the highest probability in any topic. Rows are sorted to group terms occurring within similar contexts together. [16, 17] For instance, the top rows are all tightly related to the "Systems" topic, though they are increasingly linked to the "Advanced Materials and Bio-manufacturing" topic.

3.2 Topic Model

One mechanism to answer questions such as "what topics are contained within my documents," or, "what terms are most associated with a topic" is to infer patterns from the computational distribution of the documents. Doing this without directly labeling or classifying the documents beforehand constitutes *unsupervised learning* and is called topic modeling.

Topic modeling is a group of unsupervised natural language processing methods that can identify topics in large text collections [18, 19]. These topics are weighted sets of words that imply the overall semantics of the collection and its documents. The words weights in a topic tend to represent important concepts for that topic. The overall set of topics computed for a collection of documents can serve as an interpretable decomposition that summarizes its overall content.

Latent Dirichlet Analysis (LDA) is a topic modeling approach that uses a Bag of Words representation of each document (Section 2.2.1) to infer topics from a corpus. Being a generative technique, it assumes that documents are generated from two distributions: a per-document topic distribution that assigns topic weights to documents and a per-topic word distribution that assigns word weights to topics [15]. The goal of LDA then is to recover these distributions from the data in the form of the most likely allocation of term probabilities into discrete distributions of terms, called *topics*. So, if we observe terms as collections of words within documents, using an LDA model directly assumes that each observed document is generated from some set of topics each of which have a likelihood of "emitting" any of the terms in our corpus. Estimating those probabilities and the topic mixtures that make up each document involves training on a Bag of Words representation of each document in the entire corpus simultaneously. Note that the number of topics is an important parameter that must be passed before training has begun. It determines the number of possible "types of things" any given paper can draw upon, but allows the algorithm to determine an optimal distribution of term probabilities among them.

While it is theoretically possible to determine an "optimal" number of topics given a measure of topic quality such as coherence [20], in practice this parameter should be selected carefully to aid in communication and decision making. In the case of this analysis, AMSE MSEC accepts approximately seven submission tracks per year. Perhaps as a result of this, seven topics corresponding to a seven-dimensional space was found to result in more stable results for later analytics (see below) than all other parameter values tested by the authors ($5 \leq n \leq 15$). Similarly, as LDA merely defines n distributions over terms, it is necessary for *the analyst* to interpret each distribution of terms, estimating whatever latent groupings are being detected – i.e., to *name* the topics. This is a highly subjective process, even commonly called *"reading tea leaves"* in natural language processing [21]. Naming each topic, therefore, should constitute an iterative process of design, preferably among multiple stakeholders that carefully balances interpretations with defensible data-driven justifications.

Figure 4 illustrates how a topic model allocates term probabilities from the conference to each of seven latent topics. This is trained across the entirety of the corpus (all the papers from Fig. 2). In this plot, called a "termite" visualization[5] [16], rows correspond to terms and columns to topics that have been named through an iterative process by the authors. The probability that a term is generated by a topic corresponds to the size of each circle. Rows have been sorted by the method of *spectral seriation*, such that similar rows are grouped together as much as possible. [17]

This analysis provides some interesting insights into the topics of discussion at ASME MSEC. For example, the majority of submissions to MSEC have been historically related to manufacturing *process* research. This trend can be seen in the four topics: 1) *Subtractive*, 2) *Metal Forming*, 3) *Welding*, and more recently 4) *Additive Manufacturing*. Manufacturing Systems research is also a relevant topic as shown in the *Systems* topic. The *Advanced Materials and Bio-Manufacturing* topic captures another topic of research that has recently become prevalent at the conference. Lastly, the *Optimization* topic is often cross cutting, as deals with research relevant in each of the other topics as well. This type of trend can be seen in Fig. 4 as the *Optimization* terms of importance also appear in other topics (e.g., tool is more prevalent in *Subtractive*, despite being a core term in *Optimization*).

Figure 4 provides some other interesting results allowing experts to see terms relevant across multiple topics versus terms relevant only in one topic. For example, "cloud" is highly prevalent in *Systems*, but no where else, while "=" and "mm" are cross cutting in all topics except *Systems*[6]. Terms such as "cut" or "cutting" or "chip" are most important in *Subtractive*, while "temperature" is important to both *Additive* and *Metal Forming*. While these types of insights might seem obvious to some experts, this analysis and subsequent visualization provides a quantitative look at the conference and can be used to confirm or reject inferred perspectives to further more informed discussions. It provides a quick overview of the entire conference and the topics and related terms of discussion.

Although Figure 4 only tells a *static* story, we would also like to quantify the dynamic importance of terms, similar to what was done in our keyword frequency analysis (Fig. 3) while incorporating knowledge found within a topic modeling framework. For example, "cloud" was not discussed in retrieved documents prior to 2013, as shown in Fig. 3, and yet is a major term within the *Systems* topic. Surely that topic of interest had related terms that organized around the same themes prior to the introduction of cloud computing. As such, it would be helpful to see the evolution of each *topic* over time.

3.3 Topic Term Evolution

The naïve approach to temporal topic modelling would be to partition a corpus by date before creating separate LDA models for each piece and analyzing them each individually. Unfortunately, this drastically reduces the amount of data available to train each LDA instance. In addition, the process of "reading tea leaves" implies a complete lack of consistency from year-to-year as to which topics correspond to which preceding or following year's topics, let alone global-level topics as in Fig. 4.

Instead of creating topic models for each individual year, we can instead directly model the term evolution in each topic over time. This is called Dynamic Topic Modeling, as proposed by Blei et al.[7] [22]. Analyzing terms in each topic provides the necessary context that is missing in the analysis for Fig. 3, while enabling analytics and decisions based on trends over time — a dimension missing from Fig. 4. This topic term evolution analysis can help discover the most important terms in each topic *as they age*: how they ebb and flow over time.

As opposed to the general term trend analysis in Fig. 3, Figure 5 gives insights into term usage of importance within each topic individually. For instance, an expert can identify the trends from individual terms within a topic to better discover research trends. In the topic referred to as *Advanced Materials & Biomanufacturing*, a large overarching trend in consistent use of the word "energy" in the years 2011-2016 likely reflects strong incentive in the community to investigate energy consumption and sustainability. As of 2018, the topic is now dominated by composites and fiberous materials. This indicates what would otherwise be hidden dynamics between the two, having been aligned with the same topic.

Similarly, term coupling can provide useful insights. Using "cloud" as a continuing example, the same peak occurs in 2013 as in Fig. 3, but now it is possible to analyze the other important terms that are tightly coupled within that same topic, i.e., namely, "service" and "resource". These appear to rise and fall together. However, other terms are becoming more indicative of this topic in the past few years, with "datum" and "robot" growing significantly of late.

Finally, identifying how terms are distributed across topics over time can, for example, indicate the context where a technology is most "in vogue" at a given time. "Laser" is a key defining term in the *Additive* topic until 2014-15 where laser-based *Welding* appears to be not only important but dominant in that topic. Instead, keywords like "3D printing" are more indicative of *Additive* research with lasers serving as an enabling technology,

[5] Seriation as suggested by Chuang [16] achieved via spectral method proposed by Fogel [17], as implemented in the package `Textacy` (`https://github.com/chartbeat-labs/textacy`).

[6] The term "=" was left in the analysis to provide an approximation of the importance of equations in papers, while "mm" remains to illustrate the importance of dimensions/measurements.

[7] This is accomplished by assuming that the topic distributions for documents are sampled directly from the parameters of corresponding topic distributions in the previous time slice, thus achieving smooth, contiguous topic evolution.

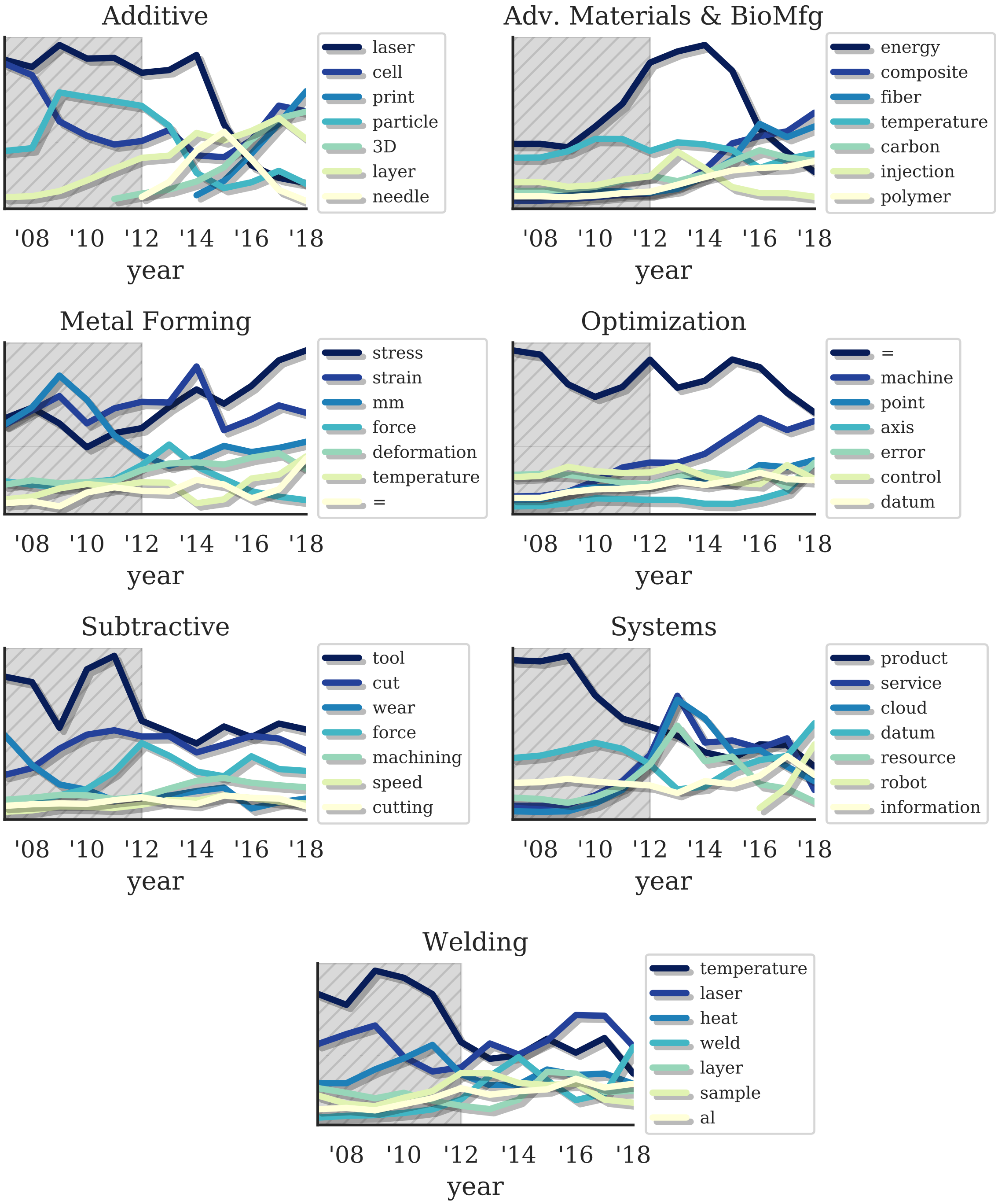


FIGURE 5. Dynamic term *importance*, i.e., relative term probabilities within dynamic topics, normalized by total topic occurrence in each year. Darker color means higher peak importance since 2012. Shaded region indicates years 2012 and before, for which downloaded document collection was incomplete.

Sexton, Thurston; Brundage, Michael; Dima, Alden A.; Sharp, Michael. "MSEC: A Quantitative Retrospective." Paper presented at MSEC: Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

These patterns let us quantify the degree to which specific topics of interest shift from one key idea to the next while maintaining an underlying connection and a more consistent relationship with the rest of the topics as a whole.

While this analysis provides a glimpse at how relevant research areas evolve over time, it is important to note that it says very little about how groups of individuals and/or papers *coalesce* around these areas through time. “Cloud” may be falling out-of-vogue within the topic overall, but ostensibly the community of researchers contributing to that body of work moved on to other relevant terms. Although terms like “robot” and “datum” are growing within their respective topic (*Systems*), this does not tell us whether they arising from the same community of interest; what are the original “cloud” researchers presenting on now? The next subsection presents a method to utilize topic modeling in conjunction with particle swarm tracking, to find and track clusters of actual documents–as opposed to topic mixtures – as they occur and migrate within the space of topics. This analysis has potential to provide a glimpse into the mixture of topics that individual communities of researchers are discussing, and how this mixture evolves over time.

3.4 Document Cluster Evolution

The idea of particle swarm tracking is not a novel one. The basic concept is to define the center of a group of similar entities (or entities that share a local region of some data space) as a swarm and track its movement through time. The number of particles that make up a group may shift over time, but so long as the local region defined by the swarm maintains a defined region and sensible movement criteria, then this local cluster can be monitored as a single entity.

As applied to this work, once a vector representation of a series of documents—i.e., conference papers—has been made, those documents can be thought to exist together somewhere in this semantic space. Using these locations, collections of proximity- or density-based clusters can be established that each contain a minimum number of documents. Labeling this group of documents as a single entity with volume and location, any number of particle tracking algorithms may then be applied to derive relative movement and interactivity of naturally occurring research efforts within the semantic document space.

In this work, a hierarchical density-based cluster approach was chosen to create the research document group instances. The specific tool used in this work is HDBSCAN from Rahman et al. [23]. This tool allowed for natural development of differing numbers of document groups, intuitively leading to new research thrust arrivals or merging or splitting of existing thrusts through time. Other clustering methods such as k-means, which specify a required number of clusters, makes such dynamic generation of topics less intuitive and forces prior assumptions upon the data.

The final step in the document group tracking is to connect the document groups to corresponding groups through time, thus making temporal traces of research thrusts. While there are many methods to accomplish this connection, for simplicity the intuitive method of requiring significant overlap across time of the document groups was selected. This method required the minimum number of assumptions and allowed for the method to be generalized to any selected characterization of the document-group space. In this work, the groups were represented as Gaussian-based N-dimensional fuzzy hyper-cubes. This allowed for rapid calculation of overlap in infinite space. With this, groups of different years can be checked for sufficient enough overlap between them to justify inclusion as a temporal trace. Although overlap was chosen as the connectivity metric here, any metric which captures significant semantic commonality between the clusters can be used to connect a cohesive trace through time.

Building each trace was performed as a single directional pass starting at 2018 and creating or extending traces one year at a time towards the beginning of the document. This single backwards stepping approach ensures that each termination point is unique for every particle trace. At each step backwards through time, the earliest point of each existing trace checks for the the document group with the most overlap (if any exist) to add as the new earliest entry. After all traces have claimed groups, any unclaimed particles are then defined as the terminal point of a new trace. Note that a single document group could be claimed by multiple traces as during their initial creation (while moving backwards in time). Subsequently, this will be shown as a split in research thrusts after construction when interpreting them forwards through time. This comparatively simple process of extending traces back through time accounts for expected behaviors of research efforts to die, merge, split, or even jump years if, for example, some new technology revitalizes some dormant research topics from the past.

There are other possibilities for creating particle traces of document clusters that may provide comparable results. The selections for this process were made to minimize the number of a priori assumptions and complex calculations to allow this process to be easily extended to higher-volume processing.

As presented in Figure 6, a total of ten different research thrusts were discovered and trended as document swarms through time. Labeled generically A through J, Fig. 6 shows that while each of the major topic areas (the colored ribbons) have one research thrust that mostly centered in it over time, there also exist several thrusts that cross cut these seven topics. Also, the focus of some of the long running thrusts seems to shift over time.

Trace A captures the primary research thrust concerning *Advanced Materials and Bio-Manufacturing*. We can see that there is a constant contribution of optimization to this research thrust, which is intuitive given that part of advanced materials research is to optimize some need via the materials or to opti-

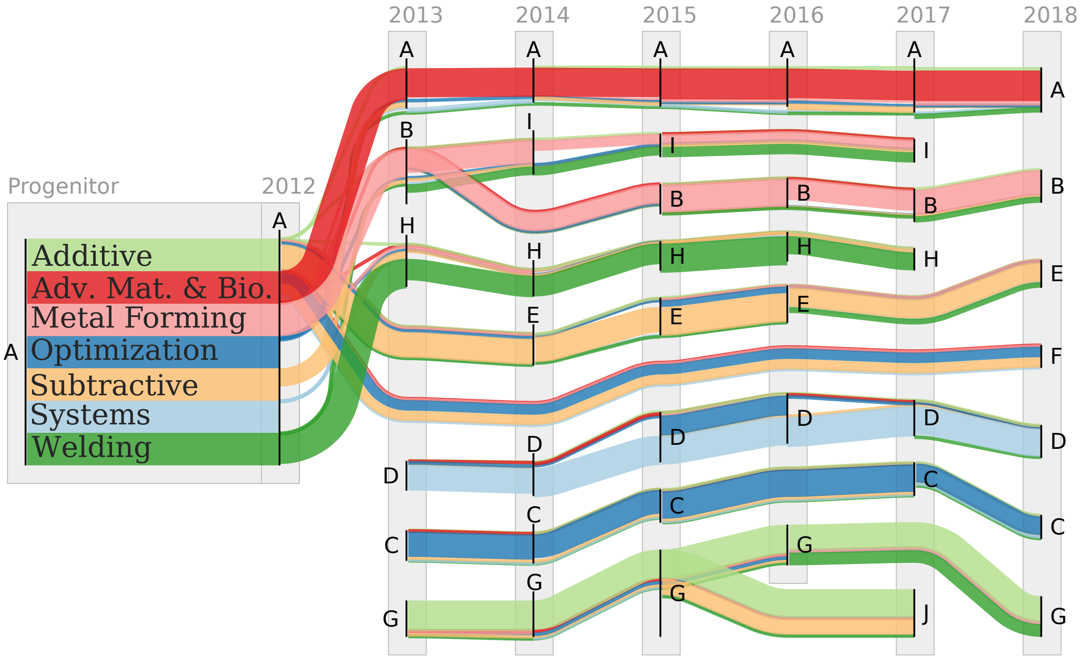


FIGURE 6. Sankey Diagram of document cluster alignment with LDA topics. Color indicates topic, while band thickness indicates proportion of that topic's global occurrence happening solely within each cluster. A trace is a presumed evolution of a persistent cluster. For example, if you look at the *Additive* topic, it is predominately in trace G until 2015 when it splits into a mixture of trace G and trace J. The key difference is the contributions from the *Subtractive* versus the *Welding* topic areas. This can be interpreted by thinking of G as a persistent research thrust that dominates the *Additive* topic until a new area splits apart form it in the form of trace J that focuses more on research related to the *Subtractive* domain while G moves into *Welding*.

mize construction methods for said material. Additionally, there are occasional spikes of interest in types of processing for this topic, namely *Subtractive* and *Welding*, which can relate again to both the use and creation of such materials. This research thrust remains the most focused through time as no other thrust has a major contribution from *Advanced Materials and Bio-Manufacturing* and no other topics contribute strongly to this thrust.

Although not as pure as A, traces B, C, D, E, H, and G each represent the central or dominate research thrust for one of the seven semantic topics. With rare exception, each of these are easy to interpret as analogs to the main topic with only slight leanings towards other topics through time. Notably, there seemed to be a brief, but strong shift towards *Optimization* in the *Systems* dominating research thrust (D) during 2015. This temporary shift may have precipitated from a shift in directives from industry drivers or some new technology hype that ultimately did not remain relevant in this research thrust community.

Some notable splits in related research can also be seen in Figure 6. For example, the divergence of I in 2014 shows that unique interest area revolving specifically around *Metal Forming* and *Welding*. This reflects a growing interest in general production methods as opposed to specific individual methods. Soon after, this area picks up a noticeable contribution from *Advanced Materials and Bio-Manufacturing*, indicating that the growing interest in use of advanced materials.

The research thrust J splits off from the *Additive* dominated thrust G in 2015 by gaining significant contributions from the *Subtractive* topic area. This research thrust would seem to again be a leaning towards more general manufacturing techniques by investigating complementary methods. Even the minor contribution from *Welding* seems to confirm this. Both major splits from the single topic dominated thrusts seems to be a step in interest towards higher-level investigations.

The only thrust area to have significant contributions from multiple topics from inception is F. Relating to mostly to *Sub-*

tractive, *Optimization*, and to a smaller degree *Advanced Materials and Bio-Manufacturing*, this thrust is likely characterizing research on the manufacturing process itself. Interest in optimizing the manufacturing process has understandably been a consistent research area in this community.

Obviously the labels and insights related to each of the thrusts are framed through the lenses of the established semantic topics. Were different numbers of semantic topics chosen, or different methods for creating these topics used, this may have lead to capturing slightly different research-thrust traces over time. With any methods of data collapse and visualization some nuance and specific information will be lost. Even so, the basic trends and results from this analysis can shed interesting insight into both trends and overall focus of research efforts.

Other revealing metrics useful for characterizing the overall research efforts in a set of documents relate to the movement and size of each of the identified document swarms. Looking at relative movement can help to estimate progress and the existence of common driving forces or technologies between efforts. On the other hand, swarm volume or counts of included documents could be used to estimate interest and participation by communities, based on the intuition that more interest will produce more publications within the trace clusters (i.e. swarms). Although omitted for space in this paper, these metrics are simple to calculate and monitor for each identified trace via the process described above.

The major pitfall of this or any particle swarm tracking method is that it must be performed with a number of entities both large enough to successfully group and distributed in such a way to characterize the behavior you wish to capture. Due to the low fraction of retrieved documents prior to 2013, any trends or insights derived are hugely unreliable during those years. As shown in Figure 6, all discovered traces (A through J) derive from a single progenitor. This is strictly due to the low number of retrieved documents for that time frame only being able to produce a single cluster. With access to more of the published documents, it is expected that multiple progenitor or initiating research points would be discovered.

4 CONCLUSIONS & FUTURE WORK

This paper discusses a methodology to determine topic trends and evolution for bodies of research publications, with a specific case study of the AMSE MSEC conference. The paper provides the steps needed to repeat the process and interpret the results in the hopes that others will use this method for quantitatively analyzing new research areas throughout time.

One important takeaway from this paper is the need for multidisciplinary teams when analyzing these results. Merging domain expertise and NLP knowledge while interpreting these topics is key when using this analysis for decision making purposes. Another important takeaway is the necessity of consulting multiple visualizations when analyzing this data; no single metric or visualization can completely describe the complex interplay of research thrusts within a community. At various times while interpreting the results for this paper, the authors simultaneously consulted multiple visualizations to get the “big” picture. For example, by looking at Fig. 3, we noticed that that the term “cloud” started to appear in 2013. We were then able to consult Fig. 4 and further analyzed it place in the topic space. Once we discovered that “cloud” appears predominantly in the *Systems* topic, we could then use Fig. 5 to analyze other important terms from that topic. Finally, Fig. 6, allowed us to analyze the evolution of clusters of papers through time.

A key improvement for this analysis pipeline would be to better facilitate the iterative, collaborative process between research domain experts and the NLP algorithms or analysts (assuming they are not the same). For instance, allowing a mixture between domain expertise-driven topic definitions and latent-topic discovery would improve interpretability and understanding for the domain experts and ensure patterns recovered by the NLP algorithms and analysts are more likely to be useful. Future work should investigate various tools for continual human-NLP collaboration, taking cues from models like Anchored Correlation Explanation [24], which allow users to guide topics toward more human-readable distributions.

An application using the method presented in this paper could be adapted to automatically predict paper placement in symposiums or to help define research sub-communities of interest within a larger community, such as a conference. Other future work could be predicting research effort movement to anticipate future interests and create topics to allow conference planners to better organize current community interests. For example, given the trends in MSEC, optimization topics with advanced materials or additive manufacturing seem to be gaining momentum and may produce a unique research thrust centered in those topics soon. Lastly, this type of analysis could be used to better predict standards needs by studying the lag time between surges in academic publications in a space (e.g., additive manufacturing) and the first mentions and development of standards in the same space. This would allow standards organizations to better adapt to high velocity technical research and start analyzing standards needs more efficiently.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

[1] ASME, 2019. *"Manufacturing Engineering Division Newsletters"*. "Available at https://community.asme.org/cfs-file.ashx/__key/communityserver-wikis-components-files/00-00-00-18-32/2158.ASME-MED-Fall-2018-Newsletter.pdf. Accessed 11-13-19".

[2] ASME, 2019. *"Manufacturing Engineering Division"*. "Available at https://community.asme.org/manufacturing_engineering_division/w/wiki/3639.about.aspx. Accessed 11-13-19".

[3] Denny, M. J., and Spirling, A., 2018. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it". *Polit. Anal.*, ***26***(2), pp. 168–189.

[4] Palmer, D. D., 2000. "Tokenisation and sentence segmentation". *Handbook of natural language processing*.

[5] Kohlschütter, C., Fankhauser, P., and Nejdl, W., 2010. "Boilerplate detection using shallow text features". In Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, ACM, pp. 441–450.

[6] Friedl, J. E. F., 2002. *Mastering Regular Expressions*, second edition ed. "O'Reilly Media, Inc.", Sebastopol, California.

[7] Manning, C. D., and Schutze, H., 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

[8] Jurafsky, D., and Martin, J., 2000. *Speech and Language Processing*. Prentice Hall, Upper Saddle River, New Jersey.

[9] Manning, C. D., Raghavan, P., and Schütze, H., 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[10] Isbell, C. L., and Viola, P., 1999. "Restructuring sparse high dimensional data for effective retrieval". In Advances in Neural Information Processing Systems, pp. 480–486.

[11] Krohn, J., Beyleveld, G., and Bassens, A., 2019. *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*. Addison-Wesley Professional, Aug.

[12] Rao, D., and McMahan, B., 2019. *Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning*. "O'Reilly Media, Inc.", Jan.

[13] Young, T., Hazarika, D., Poria, S., and Cambria, E., 2017. "Recent trends in deep learning based natural language processing".

[14] Akbik, A., Blythe, D., and Vollgraf, R., 2018. "Contextual string embeddings for sequence labeling". In Proceedings of the 27th International Conference on Computational Linguistics, pp. 1638–1649.

[15] Řehůřek, R., and Sojka, P., 2010. "Software framework for topic modelling with large corpora". In LREC 2010 workshop New Challenges for NLP Frameworks., pp. 46–50.

[16] Chuang, J., Manning, C. D., and Heer, J. "Termite: Visualization techniques for assessing textual topic models". In Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12, ACM, pp. 74–77. event-place: Capri Island, Italy.

[17] Fogel, F., d'Aspremont, A., and Vojnovic, M. "Spectral ranking using seriation".

[18] Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. "Latent dirichlet allocation". *J. Mach. Learn. Res.*, ***3***, pp. 993–1022.

[19] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M., 2009. "Reading tea leaves: How humans interpret topic models". In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, eds. Curran Associates, Inc., pp. 288–296.

[20] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A., 2011. "Optimizing semantic coherence in topic models". In Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp. 262–272.

[21] Chang, J., Boyd-Graber, J., Gerrish, S., Wang, C., and Blei, D. M. "Reading tea leaves: How humans interpret topic models". pp. 288–296.

[22] Blei, D. M., and Lafferty, J. D. "Dynamic topic models". In Proceedings of the 23rd international conference on Machine learning - ICML '06, ACM Press, pp. 113–120.

[23] Rahman, M. F., Liu, W., Suhaim, S. B., Thirumuruganathan, S., Zhang, N., and Das, G., 2016. "Hdbscan: Density based clustering over location based services". *arXiv preprint arXiv:1602.03730*.

[24] Gallagher, R. J., Reing, K., Kale, D., and Steeg, G. V. "Anchored correlation explanation: Topic modeling with minimal domain knowledge".

Application of Broadband RF Metrology to Integrated Circuit Interconnect Reliability Analyses: Monitoring Copper Interconnect Corrosion in 3D-ICs

Papa K. Amoah, Jesus Perez, and Yaw S. Obeng[#]
Nanoscale Device Characterization Division,
Physical Measurement Laboratory,
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899

Telephone: (301) 975-8093,
[#]Email: yaw.obeng@nist.gov

Abstract— **In this paper we describe the application of high-frequency electromagnetic wave (in microwave frequency / radio frequency (RF in the range, i.e., 3 kHz – 300 GHz) based techniques to probe material and structural changes that occur in integrated circuits. These techniques fall under the general area of "Broadband Dielectric Spectroscopy". In this paper, we describe corrosion of the redistribution layer (RDL), required for the implementing 3-D integrated circuits (3D-ICs), during high-temperature storage. As an illustration of our techniques, we use the RF signal loss between ports 1 and 2 on a typical vector network analyzer (i.e., RF insertion loss, S21), to monitor the oxidation of the RDL copper interconnects. We compare the RF signal loss results to the direct current-resistance that was measured simultaneously with the S21. Using electrodynamic simulations, partition the RF signal loss in corroded copper interconnects, and discuss the significance of the roughness at the air-copper oxide interface.**

I. Introduction

Recent advances in technology and materials innovations have shifted the architecture of integrated circuits from two-dimensional planar to three dimensional (3D) systems, for example, as in stacked three-dimensional integrated circuits (3D-ICs). However, the performance demands of these 3D-ICs conflict with the reliability needs of these devices. That is, the 3D-IC devices are expected to operate at higher current densities but have lower voltage tolerances at higher electric fields. Thus, with 3D-ICs, the reliability of the electronic circuitry has shifted from being transistor-dominated to interconnect-dominated. These reliability issues are mostly materials related and include such effects as resistivity change, stress-induced unexplained early failures, corrosion, electromigration, etc. The traditional metrology techniques, such as DC resistance, have been unable to adequately address the measurement needed to characterize in-situ the underlying failure mechanism.

In this paper, we describe how appropriate test structures can be used in conjunction with broadband radio frequency dielectric spectroscopy (BDS)-based metrology to fill some of these metrology gaps. Specifically, we describe how we have used BDS to study the corrosion of the redistribution layer (RDL), required for the implementing 3-D integrated circuits (3D-ICs), during high-temperature storage. The redistribution layer (RDL) allows for circuitry fan-out and allows for lateral communication between the chips [1]. Figure 1 shows the basic elements of 3D-ICs, of which the interposer is arguably the most critical. Here, we use the RF signal loss between ports 1 and 2 on a typical vector network analyzer (i.e., RF insertion loss, S21), to monitor the oxidation of the RDL copper interconnects.

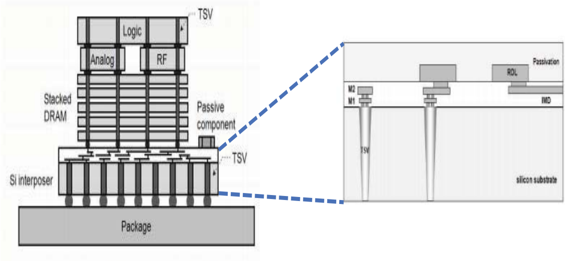


Fig. 1. Essential components of 3D stacked IC using TSV and a silicon interposer. The exploded area shows the elements within the interposer layer, pay attention to the passivation layer encapsulating the RDL (Adapted from [4])

II. Results

In a common implementation of the RDL, polymers are used as passivation and Cu-plating is used to make the metal layer. Unfortunately, in this integration scheme, the polymer passivation limits the thermal budget of the back-end of line process flow. At temperatures above the glass-transition temperature (T_{gt}) of the passivating polymers, dimensional changes expose the copper material to ambient air, resulting in spontaneous oxidation of the metal. Figure 2 compares electron micrographs of the (A) 'as-received' and (B) oxidized RDL copper metal feature. The extent of copper oxidation increases with increasing temperature and time at temperature.

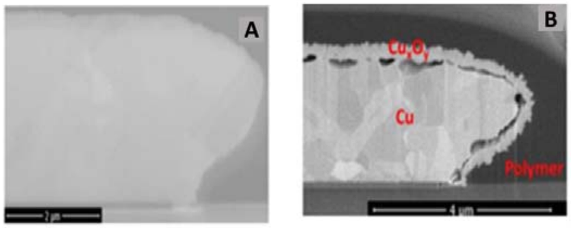


Fig. 2. Micrographs showing the development of copper oxide films around RDL feature: (A) “as-received” and (B) after 4 days at 200°C.

Figure 3 shows the microwave insertion loss (S21 at 100 MHz) as a function of the extent of corrosion of the copper RDL metallization (i.e., copper oxide film thickness measured by SEM, as in Figure 2) [2].

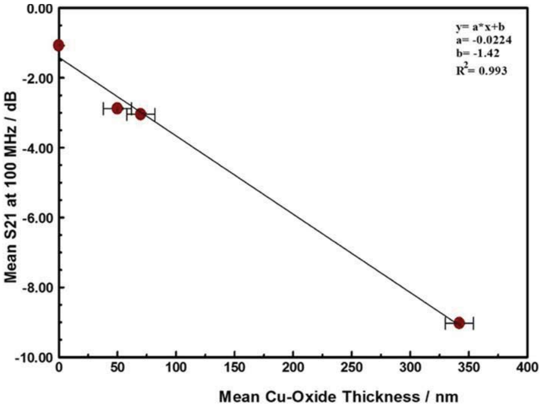


Fig. 3. Correlation between the Copper Oxide Film Thickness and the Insertion Loss (S21) at 100 MHz (Adapted from [2])

The changes in the electrical properties of the copper metal due to the oxidation are readily observed with BDS, as shown in Figure 4. In our experimental setup, we were able to monitor the direct current resistance (RDC) of the device under test (DUT). Figure 5 compares the RF signal loss results (at a single frequency (1 GHz) to the direct current-resistance that was measured simultaneously with the S21. The insertion losses increase dramatically with increasing DUT resistance (R_{DC}), suggesting that S21 may be a more sensitive measure of material changes (e.g., RDL corrosion).

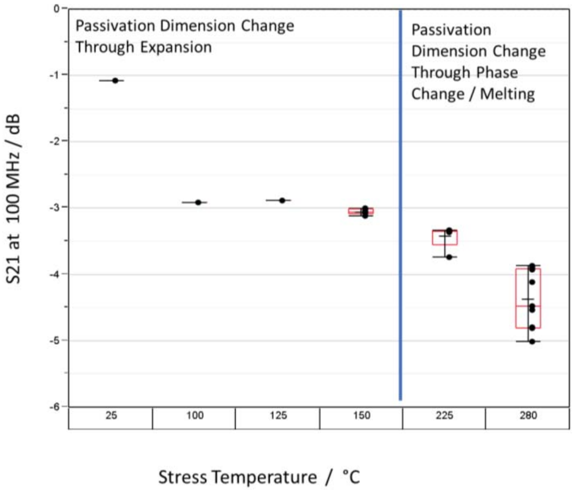


Fig 4: A comparison of the extent of oxidation (as indicated by S21) as a function of storage temperature. The data segments into two domains gated by the physico-chemical changes in the polymer passivation layer on the RDL

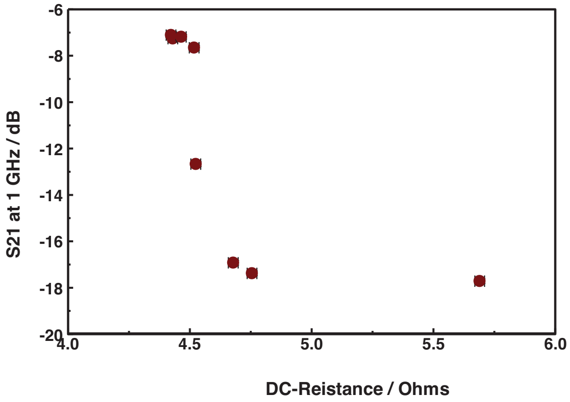


Fig 5: A comparison of the DC-resistance to the microwave insertion loss (S21) at 1GHz

Close analysis of the BDS data also provides information about the chemistry of the oxidation process, a product that techniques such as DC resistance (R_{DC}) cannot afford. For example, COMSOL Multiphysics (Burlington, MA, USA) electrodynamic modeling of the microwave signal loss (S21) through the corroded RDL required accounting for the roughness seen in Figure 2B. It turns out that the roughness of the copper oxide-air interface is attributable to the formation of electrically conductive polycrystalline copper oxide nanostructure mixtures (i.e., ranging from nanowires to nanotubes of Cu_2O and CuO) due to the thermal oxidation of crystalline copper nanowires at BEOL processing temperatures (200 to 300 °C) [3]. Figure 6, shows the simulated S21 spectrum as a function of the roughness of the oxidized RDL Cu.

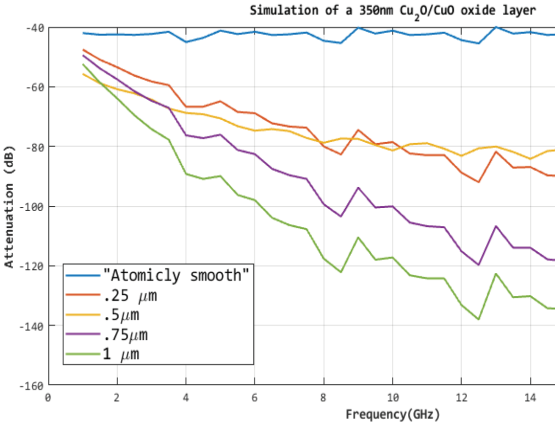


Fig. 6. Electrodynamic simulation, showing the impact of interfacial nanostructures, that manifest as roughness, on the microwave signal loss in the oxidized redistribution layer due to polymer passivation failure.

References

[1] J. H. Lau, "Recent Advances and Trends in Heterogeneous Integrations," Journal of Microelectronics and Electronic Packaging, vol. 16, pp. 45-77, 2019.

[2] P. K. Amoah, D. Veksler, C. E. Sunday, S. Moreau, D. Bouchu, and Y. S. Obeng, "Microwave Monitoring of Atmospheric Corrosion of Interconnects," ECS Journal of Solid-State Science and Technology, vol. 7, pp. N143-N149, January 1, 2018.

[3] Y.-I. Lee, Y.-S. Goo, C.-H. Chang, K.-J. Lee, N. V. Myung, and Y.-H. Choa, "Tunable Synthesis of Cuprous and Cupric Oxide Nanotubes from Electrodeposited Copper Nanowires," Journal of Nanoscience and Nanotechnology, vol. 11, pp. 1455-1458, 2011.

[4] K. Yoon, G. Kim, W. Lee, T. Song, J. Lee, H. Lee, et al., "Modeling and analysis of coupling between TSVs, metal, and RDL interconnects in TSV-based 3D IC with silicon interposer," in 2009 11th Electronics Packaging Technology Conference, 2009, pp. 702-706.

III. Discussions/Conclusions

The traditional metrology techniques, such as DC resistance, have been unable to adequately address the measurement needed to characterize in-situ the underlying failure mechanism. In this paper we have shown that the broadband radio frequency dielectric spectroscopy (BDS)-based metrology allows us to monitor the material changes in real-time, as well as extract mechanistic information from appropriately designed test structures.

Standards, Metrology and Technology to Minimize Healthcare-Associated Infections: Novel approaches to measure efficacy[#]

Brian J. Nablo[1], Darwin Reyes-Hernandez[2], Dianne L. Poster[3], Michael T. Postek[4], and Yaw S. Obeng[5]

[1]Vitreous Research Solutions, Rockville, MD
[2]Microsystems and Nanotechnology Division, Physical Measurement Laboratory, NIST, Gaithersburg, MD
[3]Material Measurement Laboratory, NIST, Gaithersburg, MD
[4]USF Health Taneja College of Pharmacy, University of South Florida, Tampa, Fl
[5]Nanoscale Device Characterization Division, Physical Measurement Laboratory, NIST, Gaithersburg, MD
* Corresponding author: yaw.obeng@nist.gov

Healthcare-associated infections (HAIs) impose great burdens on public health, making HAIs the strongest contender as the most pressing healthcare problem in acute-care hospitalization. Ultraviolet-C illumination effectively decontaminate surfaces to reduce the population of many HAI-inducing microbes. Unfortunately, no industry-accepted standards exist for evaluating the decontamination efficacy, largely due to a paucity of measurements. Ideally, technology to detect microbial populations on surfaces will be sensitive, rapid, and efficient to maintain the high turnover required by hospitals. In this paper, we discuss some recent advances in UV-antimicrobial metrology focused on underpinning the development of standards through a collaborative effort involving NIST, the ultraviolet industry and the Yale School of Medicine.

Ultraviolet-C (UV-C) irradiation decontaminate surfaces by the photodegradation of of DNA and RNA through absorption of photons resulting in formation of pyrimidine dimers from thymine and cytosine, killing a variety bacterial species, including spores [1]. The killing efficacy is most commonly quantified using basic microbiology where the number of colony-forming units (CFUs) is determined from growth plate counts of a known CFU inoculum before and after UV-C exposure. Best practices for minimizing biological and technical error in plate counting have been established [2]. Current decontamination indicators include (i) biological indicators (BI) that directly determine survivability of the most resistant microorganisms (e.g., *Geobacillus stearothermophilus*), (ii) mechanical indicators (gauges and digital displays) and (iii) chemical indicators (e.g., autoclave tape) which do not ensure sterilization but indicate that autoclaving conditions have been met. Although none of these are rapid diagnostic techniques, they are they are currently the only available metrology for evaluating UV-C decontamination and establishing room disinfection protocols [3].

Dielectric spectroscopic investigations examine and explain different molecular dynamic processes of dielectric systems in response to rapidly changing magnetic fields. In this paper, we introduce a solid-state broadband dielectric spectroscopic (BDS, 10^{-5} -10^{10} Hz) method to rapidly and nondestructively indicate decontamination of DNA containing targets on inoculated surfaces (Figure 1A), and contrast it to the UV-C induced photodecomposition of pure protein (Figure 1B). The BDS technique is predicated on detecting photoinduced changes in the electrical properties (possibly due to a reduction of the

Nablo, Brian; Reyes-Hernandez, Darwin; Poster, Dianne L.; Postek, Michael; Obeng, Yaw S. "Standards, Metrology and Technology to Minimize Healthcare-Associated Infections: Novel approaches to measure efficacy." Paper presented at Workshop on Ultraviolet Disinfection Technologies & Healthcare Associated Infections: Defining Standards and Metrology Needs, Gaithersburg, MD, US. January 14, 2020 - January 15, 2020.

cytoplasm conductance and permittivity) of DNA containing moieties [4]. Mechanistically, the observed changes in electrical properties due to UV-C exposure are the result of an eventual disruption in the semipermeable cell membrane, allowing ions and molecules to leak out of the cytoplasm, as observed with membrane-potential fluorophores in bacteria [5] and yeast [6]. The loss of cellular integrity is measurable with microwave evanescent sensors. Indeed, the changes in the electrical properties may be independent of killing method. We expect our technique to detect and quantify bacterial populations on surfaces for rapid point-of-use diagnostics, and to provide a metrology for assessing the efficacy of decontamination processes [4].

References:

[1] Nerandzic, M.M., Cadnum, J.L., Pultz, M.J. et al. Evaluation of an automated ultraviolet radiation device for decontamination of Clostridium difficile and other healthcare-associated pathogens in hospital rooms. BMC Infect Dis 10, 197 (2010) doi:10.1186/1471-2334-10-197

[2] ASTM International. D5465-93(2012) Standard Practice for Determining Microbial Colony Counts from Waters Analyzed by Plating Methods. West Conshohocken, PA; ASTM International, 2012.

[3] W. A Rutala and D. J. Wber, “Guideline for Disinfection and Sterilization in Healthcare Facilities, 2008, Update: May 2019”, Accessible version: https://www.cdc.gov/infectioncontrol/guidelines/disinfection/, accessed December 17, 2019

[4] H. Li, C. Multari, C. Palego, X. Ma, X. Du, Y. Ning, J. Buceta, J. C.M. Hwang, X. Cheng, “Differentiation of live and heat-killed E. coli by microwave impedance spectroscopy”, Sensors and Actuators B: Chemical,Volume 255, Part 2,2018, Pages 1614-1622, https://doi.org/10.1016/j.snb.2017.08.179. (http://www.sciencedirect.com/science/article/pii/S0925400517316118)

[5] S Rezaeinejad and V Ivanov, Microb Res 166 (2011), p. 129-135.

[6] J Suchodolski and A Krasowska, Microorganisms 7 (2019), p. 110

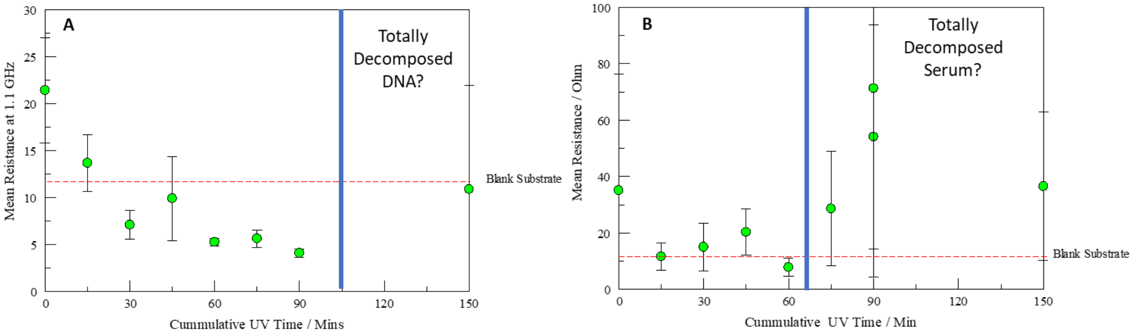


Figure 1. Side by side comparison of the evolution of the DC resistance of (A) double-stranded bacteriophage lambda and (B) fetal bovine serum (protein) thin films on glass during UV-C photolysis in open air.

Nablo, Brian; Reyes-Hernandez, Darwin; Poster, Dianne L.; Postek, Michael; Obeng, Yaw S. "Standards, Metrology and Technology to Minimize Healthcare-Associated Infections: Novel approaches to measure efficacy." Paper presented at Workshop on Ultraviolet Disinfection Technologies & Healthcare Associated Infections: Defining Standards and Metrology Needs, Gaithersburg, MD, US. January 14, 2020 - January 15, 2020.

Smart Home Security and Privacy Mitigations: Consumer Perceptions, Practices, and Challenges

Julie M. Haney[1], Susanne M. Furman[1], and Yasemin Acar[2]

[1] National Institute of Standards and Technology, Gaithersburg MD 20899, USA
{julie.haney,susanne.furman}@nist.gov
[2] Leibniz University Hannover, Germany
acar@sec.uni-hannover.de

Abstract. As smart home technology is becoming pervasive, smart home devices are increasingly being used by non-technical users who may have little understanding of the technology or how to properly mitigate privacy and security risks. To better inform security and privacy mitigation guidance for smart home devices, we interviewed 40 smart home users to discover their security and privacy concerns and mitigation strategies. Results indicated a number of concerns, but a general willingness to accept risk in lieu of perceived benefit. Concern was sometimes, but not always, accompanied by users taking mitigating actions, although most of these were simplistic and not technical in nature due to limited options or lack of user technical knowledge. Our results inform how manufacturers might empower users to take protective actions, including providing security tips and more options for controlling data being collected by devices. We also identify areas that might benefit from third-party involvement, for example by providing guidance to manufacturers on minimum privacy and security standards or developing a security and privacy rating system to aid users in selecting devices.

Keywords: smart home · internet of things · security · privacy · usability.

1 Introduction

As Internet of Things (IoT) smart home technology is becoming pervasive, smart home devices are increasingly being used by non-technical users [10] who may have little understanding of the technology or awareness of the implications of use, including considerations for privacy and security. Since their inception, smart home devices have become the target of security attacks, placing consumers' data, privacy, and safety at risk [13, 16]. In addition, concerns about the privacy and protection of potentially sensitive consumer data are surfacing [6, 12]. In fact, the U.S. Federal Bureau of Investigation (FBI) recently issued security and privacy warnings about smart televisions and other IoT devices [18, 19]. Therefore, it is critical that users are provided with the means to safeguard

Haney, Julie; Furman, Susanne M.; Acar, Yasemin. "Smart Home Security and Privacy Mitigations: Consumer Perceptions, Practices, and Challenges." Paper presented at International Conference on Human-Computer Interaction, Copenhagen, DK. July 19, 2020 - July 24, 2020.

their information and households while still enjoying the convenience of these devices.

Unfortunately, smart home device manufacturers may not provide privacy and security protections and configuration options [4], or, if they do, these options may not be transparent to the user. In addition, smart home users may not be knowledgeable enough to discern which mitigations would be most effective, or may only implement simplistic mitigations that might be inadequate [1, 26, 11, 15]. This inadequacy was demonstrated by recent stories of weak user-configured passwords being responsible for parents and children being surveilled and terrorized after their smart home devices were exploited [13].

Understanding consumers' interactions with smart home devices and their current privacy and security mitigation strategies is a first step towards developing guidance for manufacturers and third-party organizations to aid consumers. We sought to gain this understanding via an in-depth interview study of 40 smart home consumers to discover their overall experiences with, perceptions of, and challenges regarding their smart home devices. This paper addresses a subset of research questions (RQs) from the broader study that were focused on security and privacy:

RQ1: What are smart home users' privacy and security concerns, if any?

RQ2: What mitigation actions, if any, do users take to address their concerns?

RQ3: What are the factors affecting users' implementation (or lack of implementation) of privacy and security mitigations?

RQ4: What do users want (actions to take on their own or from others) in order to feel like their privacy and security are adequately protected?

We found that many users have privacy and security concerns but are mostly implementing simplistic mitigations to counter those concerns. However, some smart home users displayed a lack of concern or failed to take mitigation actions even if they do have concerns. The interviews revealed several challenges to the implementation of effective security mitigations, including users having incomplete threat models, privacy resignation, lack of transparency, poor usability of privacy and security-related device features, and lack of user technical knowledge to discern or implement appropriate mitigations. Our study makes several contributions:

– We confirm and expand upon prior studies that investigated smart home users' privacy and security concerns and mitigations [3, 20, 26] with a larger, more diverse participant sample.
– We identify several mitigations not previously described in the literature, including a more in-depth examination of smart home device updates.
– We distill participants' privacy and security "wishlist," which provides insight into potential areas for improvement in smart home device design and data handling.

– Our results inform how manufacturers and third-party evaluators might provide a more usable security and privacy experience.

2 Related Work

Prior work has examined perceptions of smart home privacy and security. Security and privacy concerns can be barriers to adoption of smart home devices. Lau et al. find that some non-users are privacy conscious and distrustful of privacy and security of smart home devices and their manufacturers, and that smart home devices generally cross these non-users' perceived privacy thresholds [11]. This finding is corroborated by Parks Associates [14], Worthy et al. [25], Emami-Naeini et al. [3], and Fruchter and Liccardy [7], who find that a lack of trust in vendors to properly safeguard personal data is a major obstacle to adoption of smart home technology. From a broader IoT perspective, Williams et al. [24] found that IoT is viewed as less privacy-respecting than non-IoT devices such as desktops, laptops, and tablets.

Adopters were found to share the same concerns, and often expressed a lack of agency in the control of their data [11]. However, they generally have higher tolerances for privacy violations, and willingly or reluctantly accept the trade-off in exchange for the convenience and utility offered by smart home devices [11]. They are generally more trusting towards well-known manufacturers and often express that they have "nothing to hide" [11, 20]. They also have complex, but incomplete threat models, which includes a general sense of being surveilled by manufacturers or the government, and the possibility of being attacked by hackers, but a lack of awareness of botnets and the sale of inferred data [1, 3, 27]. A main security concern was the possibility of a breach in the cloud that would expose user data [20].

Multiple studies discovered both technical and non-technical mitigations to address security and privacy concerns, for example passwords, secure configurations for the home network, and altering behavior around the devices [1, 11, 15, 20, 26]. However, they also identified lack of action. Reasons may be lack of awareness and availability of these options, privacy resignation, trust in the manufacturers, and assignment of responsibility to entities other than the users themselves [11, 20, 26].

Our study confirms many of the findings identified in prior literature while identifying additional mitigations such as device selection, access control, and updates. In addition, unlike other studies, we collected a wish list of mitigations that can help inform manufacturers and other entities in making privacy and security protections for smart home devices more usable for consumers.

3 Methods

We conducted semi-structured interviews of 40 smart home consumers to understand their perceptions of and experiences with smart home devices from purchase decision, to implementation, to everyday usage. The in-depth interviews

Haney, Julie; Furman, Susanne M.; Acar, Yasemin. "Smart Home Security and Privacy Mitigations: Consumer Perceptions, Practices, and Challenges." Paper presented at International Conference on Human-Computer Interaction, Copenhagen, DK. July 19, 2020 - July 24, 2020.

afforded more detailed data than could be collected via anonymous surveys and the ability to ask follow-up questions to explore responses [2]. To protect participants' confidentiality, data were recorded with generic identifiers (such as P10) and not linked back to individuals. The study was approved by the National Institute of Standards and Technology (NIST) research protections office.

We hired a consumer research company to recruit 33 general public participants, and identified seven participants via professional contacts. To determine study eligibility, adult participants interested in the study completed an online screening survey about their smart home devices, role with the devices (i.e., decision maker, purchaser, installer, administrator, troubleshooter, or user), professional background, basic demographic information, and number of household members. To ensure information-rich cases, we then purposefully selected participants who had two or more smart home devices for which they were active users.

The interview protocol addressed the following areas: understanding smart home terminology, purchase and general use, likes and dislikes, installation and troubleshooting, privacy, security, and physical safety. Interviews lasted an average of 41 minutes. Prior to the interviews, we informed the participants about the study and how we would protect their data by not recording any personal identifiers that could be linked back to the participant. All interviews were audio recorded and transcribed. General public participants were compensated with a $75 gift card.

Using widely-accepted qualitative data analysis methods [9], all three authors individually coded a subset of four interviews, then met to develop and operationalize a codebook to identify concepts within the data. Based on the codebook, we then performed iterative coding on the remainder of the interviews, with two coders per transcript. Each pair of coders met to discuss and resolve areas of difference in code application. As a group, we then progressed to the recognition of relationships among the codes and examined patterns and categories to identify themes. In this paper, we focus on themes related to privacy and security mitigations and concerns.

4 Participant Demographics

We interviewed 40 participants, 32 of whom were the installers and administrators of the devices (indicated with an A after the participant ID) and eight who were non-administrative users of the devices (indicated with a U). 55 % were male, and 45 % were female. Multiple age ranges were represented, with the majority (70 %) between the ages of 30 and 49. Overall, participants were highly educated with all but one having at least a bachelor's degree and almost half (45 %) having at a graduate degree. Table 1 shows participant demographics.

All but one participant had three or more individual smart home devices, with 34 (85 %) having three or more different types of devices. Figure 1 shows the general categories of smart home devices in participants' homes. Represented categories, along with examples of devices in that category, were:

ID	Gender	Age	Education	Occupation
P1_A	F	50-59	M	Liaison
P2_A	M	30-39	M	Lead engineer
P3_A	F	40-49	M	Professor
P4_A	M	60+	M	Retired
P6_U	F	30-39	B	Events manager
P7_A	M	30-39	B	Software engineer
P8_A	M	30-39	B	Federal employee
P9_A	F	30-39	M	Educationist
P10_A	M	30-39	B	Computer scientist
P11_A	M	50-59	M	Electrical engineer
P12_U	F	30-39	M	Administrative assistant
P13_A	M	50-59	M	Manager, Cognitive scientist
P14_U	F	40-49	H	Information specialist
P15_A	M	30-39	B	Computer scientist
P16_A	M	40-49	M	Research chief
P17_A	F	30-39	M	Systems engineer
P18_A	M	30-39	B	Business consultant
P19_A	M	50-59	B	Retail services specialist
P20_A	F	30-39	B	Administrator
P21_U	F	18-29	B	Human resources manager
P22_A	M	30-39	B	Executive admin assistant
P23_A	F	40-49	M	Community arts specialist
P24_A	M	40-49	B	Operational safety analyst
P25_A	M	30-39	B	Program management analyst
P26_A	M	30-39	B	Analyst
P27_A	F	40-49	M	Program coordinator
P28_A	F	50-59	B	Consultant
P29_A	M	18-29	M	Events coordinator
P30_U	F	18-29	B	Event planner
P31_A	F	30-39	M	Lobbyist
P32_A	M	30-39	B	Health educator
P33_A	M	18-29	B	Senior technology analyst
P34_A	M	40-49	B	Financial analyst
P35_A	M	40-49	M	Accountant
P36_A	F	30-39	B	Project manager
P37_A	F	40-49	M	Assistant principal
P38_U	F	60+	M	Special educator
P39_U	M	60+	M	Retired
P40_U	F	30-39	C	Customer service rep
P41_A	M	40-49	B	Security

Table 1. Participant Demographics. ID: A - smart home administrators/installers, U - smart home users; Education: M - Master's degree, B - Bachelor's degree, C - some college, H - High school.

Smart security: security cameras, motion detectors, door locks
Smart entertainment: smart televisions, speakers, streaming devices, other connected media systems
Home environment: smart plugs, energy monitors, lighting, smoke and air quality sensors, thermostats
Smart appliances: refrigerators, coffee pots, robot vacuums, washers
Virtual assistants: voice-controlled devices such as Amazon Echo (colloquially called Amazon Alexa) and Google Home.

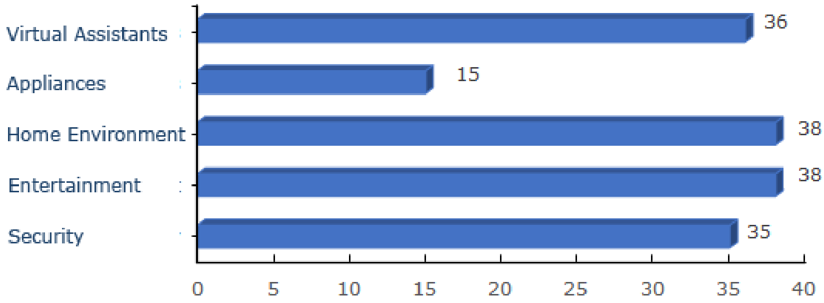


Fig. 1. Types of Smart Home devices owned by participants.

5 Results

In this section, we report results from a subset of the interview data specific to privacy and security concerns, mitigations, and mitigation wish lists. Counts of the number of participants mentioning various concepts are provided in some cases to illustrate weight or unique cases and are not an attempt to reduce our qualitative data to quantitative measures.

5.1 Concerns

We present an overview of concerns identified in our study to provide context for what our participants believe might need to be addressed by mitigations. Participants' privacy and security concerns are summarized in Table 2. For each concern in the table, we include whether the concern was discussed in a privacy or security context (or both), the number of participants mentioning each concern, and an example participant quote to illustrate the concern.

The most frequently mentioned concerns that were discussed within both the privacy and security contexts included: audio and video access via smart

	Concern	#	Example Participant Quote
Security and Privacy	Audio/video access	34	*“I was reading some article where [a virtual assistant] listens in on some of the conversations we have in our house without it being awake... That kind of freaks me out in the sense that we could be talking about something, and they have that information.” (P21_U)*
	Data breaches	17	*“Manufactures can say they can protect things, but in reality, if someone wants something bad enough, I don’t know if they really can.” (P33_A)*
	Government access	12	*“I would hate to sound like a conspiracy theorist, but I’m pretty sure the government and places like that can actually see what you do.” (P14_U)*
	Exposure of financial information	8	*“I wouldn’t want anybody committing fraud and taking my credit card information to do things they shouldn’t be doing.” (P37_A)*
Privacy	Household profiling	19	*“If someone was in control of this [device], they might be able to know what my schedule is, when I’m usually home, when the house is empty.” (P34_A)*
	Selling data	17	*“That’s what I’m really afraid of, is them packaging my information to get trends and marketing it.” (P13_A)*
	Unknowns of data collection	16	*“I’m concerned because I think we’re unaware of the types of information that these smart devices store of us or have of us.” (P21_U)*
Security	Device hacking	22	*“There’s some just people who are really smart and they’re sitting somewhere, all they’re thinking about is how to get into stuff... And if people could hack into the Department of Defense, they can hack into yours.” (P28_A)*
	Safety	17	*“It could be life threatening... If you rely on the smart device to keep your home locked,... if it does misfunction, there could be extreme circumstances. ” (P19_U)*
	Gaining Wi-Fi access	6	*“Many of these devices, you’re giving it your network password, so it has full access to everything on your network.” (P11_A)*
	Linked accounts	4	*“If you use a password commonly across different accounts, the same password, if that gets hacked... If I log into my Google account they might be able to get in because I might use the same exact password and user name.” (P2_A)*
	Poor default security settings	2	*“I would be disturbed if I saw a device that, for example, had a password you couldn’t change or restricted you to something like a 4-digit key code that’s more easily hacked.” (P15_A)*
	Update issues	2	*“I guess one area where I would be worried about would be adding features that may threaten my privacy and security.” (P15_A)*

Table 2. Smart Home Privacy and Security Concerns. # - number of participants mentioning the concern

Haney, Julie; Furman, Susanne M.; Acar, Yasemin. ”Smart Home Security and Privacy Mitigations: Consumer Perceptions, Practices, and Challenges.” Paper presented at International Conference on Human-Computer Interaction, Copenhagen, DK. July 19, 2020 - July 24, 2020.

home devices such as virtual assistants and cameras; data breaches of the manufacturer; foreign and domestic government access to data; and exposure of financial information via smart home device credentials and apps. Participants talked about the following privacy-specific concerns: household habit profiling; the selling of data and targeted ads; and unknowns about what data is being collected and how it is being used. Security-specific concerns included: general exploitation/hacking of devices; physical security/safety; gaining access to the Wi-Fi network and other devices on that network via smart home devices; gaining access to linked accounts (e.g., email or social media accounts) by exploiting device apps; poor default security settings (e.g., default passwords); and updates potentially having harmful consequences.

We also found examples of various levels of lack of concern, with seven participants having neither privacy nor security concerns. In 24 cases, participants did not think that the information collected by smart home devices was valuable or interesting to others. For example, one participant commented, *"I live a life that you could probably watch. I could probably have cameras in my house, and I wouldn't feel guilty about that... That's a concern I know some people have. But I didn't have an issue with that" (P2_A).* We also identified evidence of participants exhibiting privacy and security resignation [11, 17] (8 participants). They are of the opinion that, since so much of their data is already publicly available via other means (e.g., social media, data breaches), smart home devices pose no additional risk. One smart home user said, *"I do dislike having all of my information out there, but I think that, regardless of these smart devices, it's already out there" (P17_A).* Finally, five participants viewed exploitation of devices (hacking) as a low-probability event. This feeling was often tied to them not valuing information collected by smart home devices: *"Somebody would have to pluck us at random to really be at risk" (P25_A).*

Ultimately, even if they had concerns, participants were more than willing to accept privacy and security risks because of the perceived benefits. One participant commented, *"It's an acceptable risk if you don't think you're doing anything that's illegal or bad. It's not like I do anything weird in front of the TV besides exercise, and nobody wants to see that" (P14_U).* Another said, *"It makes my life easier, so I will continue to do it unless I have a major security concern that comes up" (P17_A).*

5.2 Mitigations

Our study discovered a variety of mitigations that participants or others in their household implement to address privacy and security concerns. All mitigations were mentioned in both the privacy and security contexts. Figure 2 shows the number of participants mentioning each mitigation. We describe the mitigations in more detail below.

Authentication. Participants mentioned using various forms of authentication (e.g., passwords, face recognition, two-factor) when asked what actions they

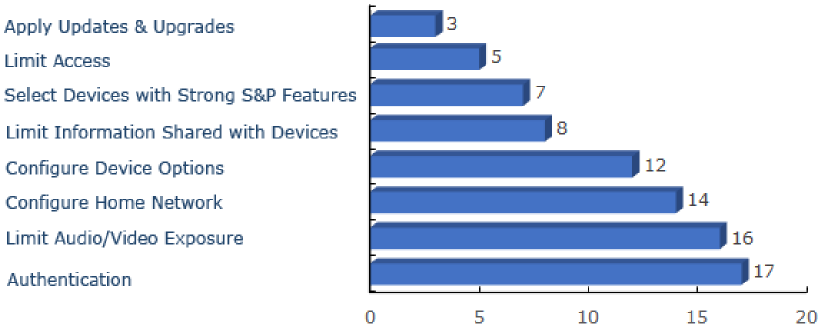


Fig. 2. Security and Privacy mitigations mentioned by participants.

take to address their concerns. However, this action was typically not a user choice, but rather prompted during installation. Authentication was most often referenced with regards to the device companion apps, which are often controlled via a cellphone.

Passwords were the most common authentication mechanism afforded by device companion apps, and often the only mitigation mentioned. One participant said that he addressed his concerns by *"password protecting the devices so nobody can connect to them... It's not very convenient, but... that's what I need to do" (P20_A)*. Several participants specifically discussed their attempts at having strong passwords: *"I have my own unique passwords that aren't dictionary words, so that's how I mitigate" (P10_A)*. Another participant used a password manager for her smart home device apps. Two others said that they made sure that they change any default passwords during installation.

Only one participant mentioned two-factor authentication in the context of mitigations: *"If I know that I can do two factor authentication for something, I'll do that" (P2_A)*. When asked about how they authenticate to their devices in a later, separate question, only one additional participant mentioned two-factor authentication, which was an option offered by his smart thermostat.

Limiting Audio and Video Exposure. To address concerns about audio and video being exposed to manufacturers or unauthorized users, study participants mostly mentioned non-technical mitigations. They were careful about where they placed cameras and virtual assistants, avoiding more private rooms in the house. For example, one participant talked about the location of his virtual assistant: *"Bedrooms are just a little more personal. I make sure not to keep it there because... if it does record, I don't want maybe those conversations and things that happened in the bedroom to be on there" (P32_A)*. Several participants were also cognizant of not having sensitive conversations in the vicinity of listening devices: *"I try to keep [my virtual assistant] in a central location and kind of avoid being close to it when having certain conversations" (P22_A).* Others covered cameras not being used. For instance, a participant remarked that her husband took action: *"The [virtual assistant] device has a video camera that you can use, but he's taped it over" (P1_A).* Finally, several users turned off devices in certain circumstances. One user talked about how her husband unplugs their virtual assistants when he is teleworking to guard against potentially sensitive conversations being recorded. Another said, *"With the security camera, sometimes I switch it off... It's when I'm really like out of town, that's when I like to switch it on back again" (P34_A).*

Network Configuration. The security and privacy of smart home devices can be contingent on the security of the home network. There were a few advanced users that mentioned more sophisticated network security mitigations, for example, segmenting their home network, installing virtual private networks (VPNs), or monitoring network traffic. For example, a do-it-yourselfer who customizes his smart home devices was diligent in securing his home network: *"I have a protective network where all these devices live in, and you can't get to it from the outside. I can get to it from within my house, and if I have to I can get to it via a VPN from the outside" (P16_A).* Another also made use of VPNs *"to mask the IP address. It's not that I'm doing anything illegal... It's just I don't feel like being tracked" (P20_A).*

However, most participants' extent of network security configuration was to password-protect their Wi-Fi. One participant commented, *"When it comes to my internet that I use to connect a lot of them, you know, it is password protected. So you know, it's not like anyone can just log on and use my network" (P32_A).* Another said, *"I'm always switching passwords with my Wi-Fi" (P34_A).*

Option Configuration. Twelve participants configured options that were at least loosely related to privacy and security. This mostly entailed disabling default functionality. For example, one participant disabled online ordering on her virtual assistant: *"We have cut off some functionality just to prevent the $400 order of mystery items" (P1_A).* A tech-savvy participant mitigated his concerns by *"turning off certain features that I think might share more information or provide more access to the device than is necessary" (P15_A)*, giving the example of how he had disabled the microphone in his smart TV. Another participant was one of the few who knew about options in virtual assistants to limit audio recording usage: *"For the [virtual assistant], it records everything. But I did see one of the options was to regularly delete it every day or something, so that kind of took the concern off the table" (P27_A).*

Limiting Shared Information. Eight participants mentioned limiting the information they share with device manufacturers, mostly when setting up companion apps. A participant said, *"I have my email address that I use for signing* *up for accounts that I'm never going to check and email address that I use for signing up for things that I actually care about. The latter is a very small number" (P17_A).* Another remarked, *"When it comes to, especially I think my [virtual assistant], I don't keep certain information stored on it. Like, I know some people will keep their actual address or even sometimes even credit card information to be able to buy things right away" (P32_A).* One participant discussed using false information when setting up her smart home device app accounts: *"I always put in fake birthdays... You need to know I'm eighteen, but you don't need to know everything" (P37_A).*

Device Selection. Some participants were proactive in their mitigation efforts by considering security and privacy in their purchase decisions. One participant remarked that, prior to selecting devices, he *"paid a lot of close attention to the security of those devices and what's happening with the data, what sorts of data they might record, how others might be able to access the system" (P15_A).* Another commented on the importance of buying secure devices: *"Even if you have to spend more money to get more into that security, we would definitely do that as we are so much dependent on this. We have to protect ourselves" (P9_A).* Others made decisions based on whether or not they trusted particular manufacturers to provide secure products. For example, a participant commented, *"I'm looking for devices that, if they're going to communicate with a cloud service, they use a well-known cloud service" (P11_A).* One made the conscious choice to buy products from well-known, larger companies: *"These are pretty big companies... We're paying money for the brand itself... Maybe that's why I'm feeling a little more secure than not... If something happens, hopefully, they have the money to figure it out" (P6_U).*

Limiting Access. Five participants made a variety of attempts to limit access to smart home devices and their apps. Three discussed limiting access of devices by visitors and service providers entering the house. One discussed making decisions on which device to use for potentially sensitive tasks, for example, *"I don't place orders via [my virtual assistant]... I do everything mostly on my computer, which has a VPN on it" (P14_U).* Another mentioned securing access to her cellphone (which contained device companion apps) as a mitigation:

> *"I'm very secure with my phone. I make sure that it's not easily accessible... I keep my phone right on me, I don't set it down, I don't let people look at stuff, I don't access the [public Wi-Fi] internet in other areas when I'm using those apps" (P37_A).*

Updates. Although updates can be a powerful mitigation against device vulnerabilities, only three participants mentioned updates or upgrades in the context of mitigations. A user said, *"I found that I'm updating everything a lot more... just kind of keeping up with the technology because it is so important" (P31_A).* A do-it-yourselfer purchased a smart camera with dubious ties to a foreign government, so he *"modified the firmware so it's no longer using the [untrusted] web service or cloud service" (P11_A).*

Prior to the security and privacy portions of the interviews, we asked participants about their experiences with device updates. Participants rarely associated updates with security or privacy and mentioned that they often do not know whether updates are available or have been installed due to inconsistent notifications and user interfaces. While updates are often viewed as potentially being security-related with traditional IT products (e.g., Microsoft's "Patch Tuesday'), we did not find that same association in our study. In addition, users often do not apply updates if they feel their devices are still working without issue. These findings indicate both a usability problem and a perception that updates are only functionality-based and not related to security.

Lack of Mitigations. We also discovered reasons for participants not implementing mitigations. Several participants cited a lack of privacy/security options or them not being aware of available options: *"Usually the description of the controls aren't specific enough... They're like, 'Check this for our privacy settings,' and sometimes the description of the settings aren't very specific" (P13_U).* Similar to reasons behind lack of concern, users often exhibited resignation and feelings of lack of control: *"I wish we could [limit data collection], but I don't think there'll ever be a way to control it" (P12_U).* Others cited a lack of knowledge or skill, especially with respect to cybersecurity: *"I'm not going to educate myself on network security... This stuff is not my forte. I'm very accepting to the fact that it is what it is" (P8_A).* Of course, some participants were simply not concerned enough to take any kind of action: *"I go on faith that they don't find me interesting enough. I guess that's it" (P23_A).*

5.3 Mitigation Wish List

Even though users have ultimately accepted privacy and security risks by introducing the devices into their homes, we found that they still desire greater control, especially with respect to privacy. We asked participants what they would like to do to protect their smart home privacy and security but are not doing, cannot do, or do not know how to do. Examination of the participant "wish list" provides insight into what would make users feel more empowered to take mitigating action and what options or instructional information they think manufacturers should provide.

Data Collection Transparency. Users desire manufacturers to be more forthcoming about what data is being collected, where it is going, and how it is being used (mentioned by 12 participants). Manufacturers claim that user level agreements provide this information. However, participants said that they rarely read the long agreements and generally do not find those useful because they are in *"lawyer speak. You don't really know what they're collecting because they can use* *language to mislead you" (P31_U).* The lack of transparency leaves users wanting more: *"At least give us notice in terms of who has access to it... We would appreciate that and make us feel more comfortable around the security behind it" (P21_U).* One user desired a more concise, clear statement of data usage: *"if these companies provided a manifesto of what information they're interested in or how they use information and how they're collecting information and provide that - a one pager - that would be great" (P2_A).* Realizing that it might not be in manufacturers' best interest to clearly disclose data usage, P31_A saw the government as having a role since *"we've got to do something to protect people's information, or at least make them more aware of what exactly is being utilized and sold."*

Privacy and Security Controls. Ten participants would like more control over the devices and data. This includes the ability to opt in/out of various data collections, limit how data is shared, and configure security and other privacy options. For example, a participant remarked, *"there would be some of these products that I have been avoiding purchasing that I might purchase if they provided more granular control over... all aspects of the security and privacy" (P15_A).* Another participant said he would like to be able to use two-factor authentication for his devices' companion apps: *"There would be features that would be nice to have, I guess one being a two-factor authentication. If my phone is close to my thermostat, that's my second factor" (P10_A).* Options should also be easy-to-configure, as mentioned by one participant: *"I think the ability to control that data should be simpler than a multi-step process" (P29_A).*

Technically advanced users were more specific about what they would like to do and wanted granular controls. A computer scientist said, *"I would really be happy actually if a lot of them had APIs [application programming interfaces] that I could use to directly program their behavior and get more control over them programmatically" (P15_A).* An electrical engineer commented:

> *"I'd like to have the ability to potentially allow or disallow the functionality of all these devices, maybe at given times. I'd like to be able to define what are allowable communications or protocols" (P11_A).*

Five participants wished that they had the ability to keep smart home data on their local network when possible instead of the common business model of data being sent to manufacturers or their cloud services. A participant said, *"If I could not have accounts and just have it on my own home network, I would prefer that" (P17_A).* P15_A commented that he wished *"some of these devices used the voice control features locally only rather than sending clips of your voice over the Internet to be analyzed."*

Security Feature Transparency. Four participants would like to know the level of security provided by the devices. One stated, *"it would be nice to know what security features are already there because they're not advertised or transparent at all. And maybe to have an option to get some kind of enhanced security if you wanted to" (P24_A).* Wishing to know if he needed to bolster the security of his home network to counter potentially weak smart home security, another participant said, *"I wish I knew more about what kind of encryption they use" (P3_A).*

Assistance for Users. Within the security context, four participants expressed their desire to be provided with suggestions and instructions on how to better secure their devices. A participant unfamiliar with security best practices commented, *"I think I need to be advised on good practices that I could take... And then I probably would implement them" (P35_A).* Another suggested, *"maybe the apps that I have could throw out reminders in a more frequent manner that says are you doing something like this to protect yourself?" (P19_A).* A heavy user of smart home devices said that he would like to know how best to protect his devices against vulnerabilities: *"I would like the vulnerability identified well enough so I know what it is and then some directions on how to solve it" (P13_A).*

6 Implications

The users we interviewed were diverse in their mitigation approaches to smart home devices. Some were proactive from a privacy and security perspective and knowledgeable about the technology. Others had very little understanding of the technology and implications of use. Our results suggest that users do the best they can with the skills and the options available to them.

Most of the mitigations identified in our study were simplistic (e.g., setting passwords) or not technical in nature (e.g., placement of devices). From a privacy perspective, participants expressed the desire to be able to control what happens to their data but do not know what options are available, or, in many cases, no options exist. Security concepts and implications were more difficult for participants to grasp, with many lacking the knowledge to implement effective mitigations, for example, by properly securing their home networks. Overall, we observed that many of the participants were left with a feeling of discomfort because they had privacy and security concerns but felt powerless to address those.

Based on study results, we describe possible ways in which manufacturers could empower users to make appropriate security choices through usable interfaces and where further research may be helpful. We also identify areas that could benefit from third-party evaluation and guidance.

6.1 Considerations for Usable Security and Privacy Options

Participants' current mitigation strategies (or lack thereof) and their wish lists for privacy and security can inform what additional options manufacturers could provide and other areas where they might alleviate user burden by defaulting to strong privacy and security.

Note that since our interview study was broader than privacy and security, we had the opportunity to delve into users' installation and administration experiences with their smart home devices. Participants revealed that they rarely change settings after initial setup. Therefore, additional research may be warranted to investigate if installation is the best time to prompt users on security and privacy options.

Secure and private by default: As revealed in prior usable security research, people are often reluctant to change default security settings [28, 29]. Therefore, to alleviate undue burden on users, there may be settings which manufacturers could configure to be the most secure/private by default. However, more research should be conducted to understand how setting defaults to the most secure/private options may contribute to or detract from usability.

Opt in/out: Currently, opting out of data collection and various uses may not be possible or may be burdensome. For example, P17_A said that one manufacturer required a letter be mailed requesting to limit data sharing. Based on participants desiring more control on data usage, more research is needed regarding how manufacturers could offer easy-to-configure opt in/out options.

Data usage transparency: Device privacy policies and user agreements are rarely read and difficult to understand, leaving users uninformed about data collection practices. Manufacturers could provide greater transparency about what data is collected, where the data goes, how long it is stored, and who it is shared with.

Data localization: Our participants were often concerned about manufacturer profiling of their households, selling of their data, and possible data breaches of manufacturer data storage. To counter these concerns, manufacturers could provide options to localize whatever data processing can be localized instead of sending everything to the manufacturer's cloud.

Securability: In situations where security settings might be dependent on user context, there could be a focus on "securability," which is the "ability and knowledge to enable and configure the appropriate security features" [23]. To achieve product securability, manufacturers could facilitate secure use by providing users with real-time assistance, such as configuration wizards, to help them set the level of security appropriate for their situation. For example, users might be given the option of configuring low, medium, and high levels of security based on clear criteria (e.g., network environment, context of use, risk tolerance) gleaned through a security configuration wizard. The securability concept can also be applied to privacy settings.

Granular options for advanced users: We interviewed several advanced users who were well-versed in technology and security. These users wanted more control over security settings. Therefore, in addition to supporting less technical users with guided wizards and instructions, manufacturers could offer more granular security controls for those who want them. We acknowledge that striking the right balance between an abundance of granular options and a minimal set for less-technical users may be difficult. Therefore, we recommend additional research into interface solutions that may attempt to balance these considerations.

Update transparency: Updates are especially important as they might be the only mitigations for certain kinds of smart home device vulnerabilities (e.g., those in the code). In line with the NIST Interagency Report 8267 (Draft) Security Review of Consumer Home Internet of Things (IoT) Products [5] recommendation that users receive update notifications in a timely manner, manufacturers might either provide an option for automatic updates or push notifications to users with clear installation instructions and descriptions of the importance of applying the update.

Network security tips: Home networks need to be secured to protect smart home devices. However, people often lack the knowledge and motivation to take action. For example, the FBI recommends that users segment their network [13] even though few participants in our study had the technical knowledge to be able to do so. Several of our study participants said they would like manufacturers to provide step-by-step tips on home network security (e.g., setting up secure Wi-Fi, password-protecting all devices on the network) that complement the security options provided by the devices themselves.

6.2 Third-party Opportunities

Our results suggest that users may be open to third-party organizations (e.g., government agencies, industry groups, standards organizations) playing a bigger role in suggesting guidance for manufacturers concerning the usability of smart home security and privacy features and options. For example, the guidance produced by NIST [4, 5] provides recommendations but emphasizes that these should be tailored to specific contexts of use while not placing undue burden on the user.

The wide variety of mitigations mentioned by participants may also indicate a need for more standardization of privacy and security best practices for smart home users by trusted third parties (e.g., government agencies or an IoT industry consortium). To help users understand privacy and security implications of smart home devices, we also recommend exploring the usability considerations of having an independent, third-party ratings system similar to that which has been proposed by the Canadian Internet Society [21] and the U.S. Government Departments of Commerce and Homeland Security [22]. This ratings system would help consumers to make informed decisions about which devices to bring into their homes.

7 Limitations

In addition to typical limitations of interview studies (e.g., recall, self-report, and social desirability biases), our study may be limited in generalizability. The small sample of participants, the majority of whom were well-educated individuals living in a high-income metropolitan area, may not be fully representative of the U.S. smart home user population. However, our study population appears to mirror early adopters of smart home devices, which have been characterized in prior industry surveys [8]. We also recognize that smart home users in the U.S. may have different privacy and security attitudes from users in other countries because of political or cultural factors, for example those related to privacy expectations. Finally, our study does not capture perceptions of those choosing not to adopt smart home technologies or limited adopters (those with only one device). Non-adopters' and limited adopters' perceptions of privacy and security could shed light on additional areas needing improvement. However, even given the limitations, our exploratory study is a solid step in investigating smart home users' perceptions and practices and can inform subsequent surveys of broader populations, for example via quantitative surveys distributed in multiple countries.

8 Conclusion

We interviewed 40 smart home users to discover their security and privacy concerns and mitigation strategies. Results indicated a number of concerns, but a willingness to accept risk in exchange for perceived benefit. Concern was sometimes, but not always, accompanied by users taking mitigating actions, although most of these actions were simplistic due to limited options or lack of user technical knowledge.

Improving the security and privacy of smart home devices will be critical as adoption of these technologies increase. Efforts should be joint between consumers, manufacturers, and third-party organizations with special consideration made for designing usable interfaces that empower users to take protective actions while not overburdening them.

Disclaimer

Certain commercial companies or products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the companies or products identified are necessarily the best available for the purpose.

References

1. Abdi, N., Ramokapane, K.M., Such, J.M.: More than smart speakers: Security and privacy perceptions of smart home personal assistants. In: Proceedings of the Fifteenth Symposium on Usable Privacy and Security (2019)
2. Corbin, J., Strauss, A.: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Sage Publications, Thousand Oaks, CA, 4th edn. (2015)
3. Emami-Naeini, P., Dixon, H., Agarwal, Y., Cranor, L.F.: Exploring how privacy and security factor into IoT device purchase behavior. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM (2019)
4. Fagan, M., Megas, K.N., Scarfone, K., Smith, M.: Second Draft NISTIR 8259 Foundational activities and core cybersecurity device capability baseline for IoT manufacturers (2020), https://doi.org/10.6028/NIST.IR.8259-draft
5. Fagan, M., Yang, M., Tan, A., Randolph, L., Scarfone, K.: Draft NISTIR 8267 Security review of consumer home internet of things (IoT) products (2019), https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8267-draft.pdf
6. Federal Trade Commission: VIZIO to pay $2.2 million to FTC, State of New Jersey to settle charges it collected viewing histories on 11 million smart televisions without users' consent (2017), https://www.ftc.gov/news-events/press-releases/2017/02/vizio-pay-22-million-ftc-state-new-jersey-settle-charges-it
7. Fruchter, N., Liccardi, I.: Consumer attitudes towards privacy and security in home assistants. In: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. ACM (2018)
8. GfK: Future of smart home study global report (2016), https://www.gfk.com
9. Glaser, B.G., Strauss, A.L.: Discovery of grounded theory: Strategies for qualitative research. Routledge (2017)
10. GutCheck: Smart home device adoption (2018), https://resource.gutcheckit.com/smart-home-device-adoption-au-ty
11. Lau, J., Zimmerman, B., Schaub, F.: Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. In: Proceedings of the ACM on Human-Computer Interaction. ACM (2018)
12. Lee, T.B.: Amazon admits that employees review small "sample" of Alexa audio (April 2019), https://arstechnica.com/tech-policy/2019/04/amazon-admits-that-employees-review-small-sample-of-alexa-audio/
13. Murdock, J.: Ring security cameras pose a threat to families and the public, privacy campaigners claim amid surge in hack attacks (December 2019), https://www.newsweek.com/amazon-ring-camera-hacking-privacy-groups-fight-future-threat-families-public-1477709
14. Parks Associates: State of the market: Smart home and connected entertainment (2019), http://www.parksassociates.com/bento/shop/whitepapers/files/ParksAssoc-OpenHouseOverview2018.pdf
15. PwC: Smart home, seamless life (January 2017), https://www.pwc.fr/fr/assets/files/pdf/2017/01/pwc-consumer-intelligence-series-iot-connected-home.pdf
16. Security Research Labs: Smart spies: Alexa and google home expose users to vishing and eavesdropping (2019), https://srlabs.de/bites/smart-spies/
17. Stanton, B., Theofanos, M.F., Prettyman, S.S., Furman, S.: Security fatigue. IT Professional **18**(5), 26–32 (2016)

18. Steele, B.A.: Oregon FBI Tech Tuesday: Securing Smart TVs (November 2019), https://www.fbi.gov/contact-us/field-offices/portland/news/press-releases/tech-tuesdaysmart-tvs
19. Steele, B.A.: Tech Tuesday: Internet of things (IoT) (December 2019), https://www.fbi.gov/contact-us/field-offices/portland/news/press-releases/tech-tuesday-internet-of-things-iot
20. Tabassum, M., Kosinski, T., Lipford, H.R.: "I don't own the data": End user perceptions of smart home device data practices and risks. In: Fifteenth Symposium on Usable Privacy and Security (2019)
21. The Internet Society: Securing the internet of things: A Canadian multistakeholder process draft report (2019), https://iotsecurity2018.ca/wp-content/uploads/2019/02/Enhancing-IoT-Security-Draft-Outcomes-Report.pdf
22. U.S. Departments of Commerce and Homeland Security: A report to the president on enhancing the resilience of the internet and communications ecosystem against botnets and other automated, distributed threats (May 2018), https://www.commerce.gov/page/report-president-enhancing-resilience-against-botnets
23. Vasserman, E., Fitzgerald, B.: Cyber-"securability" (September 2019), presentation at FDA Science Forum
24. Williams, M., Nurse, J.R., Creese, S.: Privacy is the boring bit: User perceptions and behaviour in the internet-of-things. In: 15th Annual Conference on Privacy, Security and Trust. pp. 181–18109. IEEE (2017)
25. Worthy, P., Matthews, B., Viller, S.: Trust me: doubts and concerns living with the internet of things. In: ACM Conference on Designing Interactive Systems. pp. 427–434. ACM (2016)
26. Zeng, E., Mare, S., Roesner, F.: End user security and privacy concerns with smart homes. In: Thirteenth Symposium on Usable Privacy and Security (2017)
27. Zheng, S., Apthorpe, N., Chetty, M., Feamster, N.: User perceptions of smart home IoT privacy. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW) (2018)
28. Zurko, M.E.: User-centered security: Stepping up to the grand challenge. In: Proceedings of the 21st Annual Computer Security Applications Conference. p. 14 (2005)
29. Zurko, M.E., Kaufman, C., Spanbauer, K., Bassett, C.: Did you ever have to make up your mind? What Notes users do when faced with a security decision. In: Proceedings of the 18th Annual Computer Security Applications Conference. pp. 371–381 (2002)

QUANTIFICATION OF SEISMIC ENERGY DEMAND IN NONLINEAR STRUCTURES THROUGH ANALYTICAL DERIVATIONS

S.L. McCabe(1), K.K.F. Wong(2)

(1) *Director, National Earthquake Hazards Reduction Program, USA, smccabe@nist.gov*
(2) *Research Structural Engineer, National Institute of Standards and Technology, USA, kfwong@nist.gov*

Abstract

The classical seismic design procedure uses a force-based approach where reduced seismic force demands are evaluated and compared to the corresponding seismic capacities. An improved performance-based design procedure recently developed uses the displacement-based approach to ensure that structures with nonlinear deformations perform within acceptable limits. However, neither of these procedures considers the effect of cyclic behavior that is typically observed in the seismic response of nonlinear structures.

On the other hand, energy is the product of force and displacement, which captures both monotonic and cyclic behavior of structural response. At the same time, structures suffering damage from a major earthquake participate in an energy transfer process. Earthquake ground motion transfers energy to individual structures as input energy that induces vibrations in the structure and its contents, resulting in response in terms of potential energy, kinetic energy, and damping energy. Quantifying these energy demands is often difficult, particularly in the calculation of potential energy, where earthquake often causes both material and geometric nonlinearities in the structure. Since the potential energy relates to the stiffness of the structure, it consists of three parts: (1) Strain energy associated with the elastic portion of the material; (2) Higher-order energy associated with geometric nonlinearity of the structure; and (3) Plastic energy associated with material nonlinearity of the structural components.

In this research, an analytical derivation is used to quantify the energy demand in fully nonlinear framed structures using an energy balance approach. A moment-resisting steel frame is used to verify the method by evaluating the energy response and the transfer among different energy forms throughout the dynamic analysis. The result shows that plastic energy dissipated at each plastic hinge is a positive, scalar quantity that can be calculated uniquely, and the sum of individual plastic energy at each plastic hinge is exactly equal to the overall plastic energy of the structure. Once energy demand is quantified uniquely, structural performance and damage can be assessed more easily via comparison of the energy demand with the corresponding energy capacity. Through this process of quantifying the seismic performance of structures, higher level of confidence can be achieved in the design over the current force-based or displacement-based methods.

Keywords: Energy demand; Energy capacity; Plastic energy; Material nonlinearity; Geometric nonlinearity

Wong, Kevin K F; McCabe, Steven. "Quantification of Seismic Energy Demand in Nonlinear Structures Through Analytical Derivations."
Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Current seismic design procedures use either a force-based or displacement-based approach to ensure that structures can perform adequately during major events. However, neither of these procedures considers the effect of cyclic behavior that is typically observed in the seismic response of nonlinear structures. Accumulation of damage as a function of the number of inelastic cycles has been recognized in laboratory as well as post-earthquake reconnaissance. The question often raised is how to relate maximum responses with some measure of damage. Recognizing this shortcoming in the current procedure, energy, as a product of force and displacement, can capture both monotonic and cyclic behavior. Therefore, a study on how energy can be used in structural design is worthwhile, particularly how energy demand and capacity are quantified.

Research on seismic energy began in the 1980s when it was recognized that significant cumulative damage can occur in structures without large global displacement responses in long-duration earthquake ground motions [1-4]. However, these studies focused on evaluating the hysteretic energy by calculating the enclosed area in a force-deformation curve of single degree of freedom systems. In 1990, an analytical procedure was developed to clearly define different forms of energy for linear multi-degree of freedom structures [5], and this procedure is later extended in 2002 to define energy due to inelastic deformation [6]. However, these studies did not consider the reduction of lateral stiffness by axial load due to geometric nonlinearity, which can lead to considerable error in the energy calculations.

In this research, an analytical method for calculating energy considering both geometric and material nonlinearities is derived to investigate the energy of framed structures responding nonlinearly to earthquakes. In particular, the potential energy that is directly related to nonlinear stiffness of the structure is investigated, and a step-by-step analysis of a one-story frame is used to verify that input energy is balanced by potential energy throughout the static loading and unloading phases. This potential energy consists of three parts in a fully nonlinear system: (1) the stored linear elastic "strain energy"; (2) the "higher-order energy" associated with geometric nonlinearity; and (3) the "plastic energy" dissipated by material nonlinearity. Finally, four-story frame is used to verify energy balance can also be achieved in a dynamic context.

2. Inelastic Displacement for Material Nonlinearity

The derivation on the use of inelastic displacement for analyzing structures with material nonlinearity has previously been published [6-7] and is briefly summarized here with an extension to include geometric nonlinearity. Consider a framed structure having a total of n degrees of freedom (DOFs) and m plastic hinge locations (PHLs). Let the $n \times 1$ total displacement $\mathbf{x}(t)$ at each DOF be represented as the summation of the $n \times 1$ elastic displacement $\mathbf{x}'(t)$ and the $n \times 1$ inelastic displacement $\mathbf{x}''(t)$:

$$\mathbf{x}(t) = \mathbf{x}'(t) + \mathbf{x}''(t) \tag{1}$$

Similarly, let the $m \times 1$ total moment $\mathbf{m}(t)$ at the PHLs of a moment-resisting frame be separated into the $m \times 1$ elastic moment $\mathbf{m}'(t)$ and the $m \times 1$ inelastic moment $\mathbf{m}''(t)$:

$$\mathbf{m}(t) = \mathbf{m}'(t) + \mathbf{m}''(t) \tag{2}$$

This inelastic moment $\mathbf{m}''(t)$ is often known as the "residual moment" that is caused by material nonlinearity in the structure. The displacements in Eq. (1) and the moments in Eq. (2) are related by the equations:

$$\mathbf{m}'(t) = \mathbf{K}'(t)^T \mathbf{x}'(t) \quad , \quad \mathbf{m}''(t) = -\left[\mathbf{K}''(t) - \mathbf{K}'(t)^T \mathbf{K}(t)^{-1} \mathbf{K}'(t)\right] \mathbf{\Theta}''(t) \tag{3}$$

where $\mathbf{\Theta}''(t)$ is the $m \times 1$ plastic rotation at each PHL, and $\mathbf{K}(t)$, $\mathbf{K}'(t)$, and $\mathbf{K}''(t)$ are time-varying stiffness matrices due to geometric nonlinearity in which columns are subjected to time-varying axial compressive forces $P(t)$. Since the total axial force in all columns within a single floor remains constant, software developers generally assume that all stiffness matrices remain constant throughout the duration of the earthquakes to simplify the calculations without losing accuracy. Doing so, Eq. (3) becomes

$$\mathbf{m}'(t)=\mathbf{K}'^{T}\mathbf{x}'(t) \quad , \quad \mathbf{m}''(t)=-\left[\mathbf{K}''-\mathbf{K}'^{T}\mathbf{K}^{-1}\mathbf{K}'\right]\boldsymbol{\Theta}''(t) \tag{4}$$

The relationship between the plastic rotation $\boldsymbol{\Theta}''(t)$ and inelastic displacement $\mathbf{x}''(t)$ can be written as

$$\mathbf{x}''(t)=\mathbf{K}^{-1}\mathbf{K}'\boldsymbol{\Theta}''(t) \tag{5}$$

Substituting both equations in Eq. (4) into Eq. (2) and then making use of Eqs. (1) and (5) gives the governing equation for calculating the plastic hinge responses for any total displacement pattern $\mathbf{x}(t)$:

$$\mathbf{m}(t)+\mathbf{K}''\boldsymbol{\Theta}''(t)=\mathbf{K}'^{T}\mathbf{x}(t) \tag{6}$$

3. Dynamic Equilibrium Equation of Motion

For a moment-resisting framed structure modeled as an n DOF system and subjected to earthquake ground motions, the dynamic equilibrium equation of motion can be written as

$$\mathbf{M}\ddot{\mathbf{x}}(t)+\mathbf{C}\dot{\mathbf{x}}(t)+\mathbf{K}\mathbf{x}'(t)=-\mathbf{M}\ddot{\mathbf{g}}(t)-\mathbf{F}_a(t) \tag{7}$$

where $\mathbf{M}$ is the $n\times n$ mass matrix, $\mathbf{C}$ is the $n\times n$ damping matrix, $\dot{\mathbf{x}}(t)$ is the $n\times 1$ velocity vector, $\ddot{\mathbf{x}}(t)$ is the $n\times 1$ acceleration vector, $\mathbf{K}$ is the same $n\times n$ stiffness matrix presented in Eq. (4), $\ddot{\mathbf{g}}(t)$ is the $n\times 1$ earthquake ground acceleration vector corresponding to the effect of ground motion at each DOF, and $\mathbf{F}_a(t)$ is the $n\times 1$ vector of additional forces imposed on the structure due to geometric nonlinearity accounting for all the gravity columns in the structure (mainly the P-Δ effect). In a two-dimensional analysis, the relationship between this lateral force $\mathbf{F}_a(t)$ and the lateral displacement $\mathbf{x}(t)$ can be written as

$$\mathbf{F}_a(t)=\mathbf{K}_a\mathbf{x}(t) \tag{8}$$

where $\mathbf{K}_a$ is an $n\times n$ stiffness matrix that is a function of the gravity loads on the leaning column and the corresponding story height, which can be written in the form:

$$\mathbf{K}_a=\begin{bmatrix} -Q_1/h_1-Q_2/h_2 & Q_2/h_2 & 0 & \cdots & 0 \\ Q_2/h_2 & -Q_2/h_2-Q_3/h_3 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & Q_{n-1}/h_{n-1} & 0 \\ \vdots & \ddots & Q_{n-1}/h_{n-1} & -Q_{n-1}/h_{n-1}-Q_n/h_n & Q_n/h_n \\ 0 & \cdots & 0 & Q_n/h_n & -Q_n/h_n \end{bmatrix} \tag{9}$$

Here, Q_i is the total axial force due to gravity loads acting on the leaning column of the i^{th} floor, and h_i is the story height of the i^{th} floor. Now substituting Eq. (8) into Eq. (7) and rearranging terms gives

$$\mathbf{M}\ddot{\mathbf{x}}(t)+\mathbf{C}\dot{\mathbf{x}}(t)+\mathbf{K}\mathbf{x}'(t)+\mathbf{K}_a\mathbf{x}(t)=-\mathbf{M}\ddot{\mathbf{g}}(t) \tag{10}$$

Since $\ddot{\mathbf{y}}(t)=\ddot{\mathbf{x}}(t)+\ddot{\mathbf{g}}(t)$ where $\ddot{\mathbf{y}}(t)$ is the $n\times 1$ absolute acceleration vector, substituting this equation into Eq. (10) gives the governing equation of motion for energy balance:

$$\mathbf{M}\ddot{\mathbf{y}}(t)+\mathbf{C}\dot{\mathbf{x}}(t)+\mathbf{K}\mathbf{x}'(t)+\mathbf{K}_a\mathbf{x}(t)=\mathbf{0} \tag{11}$$

4. Energy Balance

To evaluate seismic energy, integrating Eq. (11) over the path of displacement response gives

$$\int_0^t \ddot{\mathbf{y}}(t)^T\mathbf{M}d\mathbf{x}+\int_0^t \dot{\mathbf{x}}(t)^T\mathbf{C}d\mathbf{x}+\int_0^t \mathbf{x}'(t)^T\mathbf{K}d\mathbf{x}+\int_0^t \mathbf{x}(t)^T\mathbf{K}_a d\mathbf{x}=0 \tag{12}$$

Note that $d\mathbf{x}(t)=d\mathbf{y}(t)-d\mathbf{g}(t)$, where $\mathbf{y}(t)$ is the $n\times 1$ absolute displacement vector and $\mathbf{g}(t)$ is the $n\times 1$ earthquake ground displacement vector. Substituting this equation into the first integral of Eq. (12) gives

$$\int_0^t \ddot{\mathbf{y}}(t)^T \mathbf{M} d\mathbf{y} + \int_0^t \dot{\mathbf{x}}(t)^T \mathbf{C} d\mathbf{x} + \int_0^t \mathbf{x}'(t)^T \mathbf{K} d\mathbf{x} + \int_0^t \mathbf{x}(t)^T \mathbf{K}_a d\mathbf{x} = \int_0^t \ddot{\mathbf{y}}(t)^T \mathbf{M} d\mathbf{g} \tag{13}$$

In addition, the incremental displacement $d\mathbf{x}(t)$ can be separated into elastic displacement and inelastic displacement as $d\mathbf{x}(t) = d\mathbf{x}'(t) + d\mathbf{x}''(t)$. Substituting this equation into the third integral of Eq. (13) gives

$$\int_0^t \ddot{\mathbf{y}}(t)^T \mathbf{M} d\mathbf{y} + \int_0^t \dot{\mathbf{x}}(t)^T \mathbf{C} d\mathbf{x} + \int_0^t \mathbf{x}'(t)^T \mathbf{K} d\mathbf{x}' + \int_0^t \mathbf{x}'(t)^T \mathbf{K} d\mathbf{x}'' + \int_0^t \mathbf{x}(t)^T \mathbf{K}_a d\mathbf{x} = \int_0^t \ddot{\mathbf{y}}(t)^T \mathbf{M} d\mathbf{g} \tag{14}$$

Each integral in Eq. (14) is considered separately in the following sub-sections.

4.1 Kinetic Energy (*KE*)

The first integral on the left-hand side of Eq. (14) represents the absolute kinetic energy (*KE*) and can be evaluated using the absolute velocity of the structure as:

$$KE(t_k) = \int_0^{t_k} \ddot{\mathbf{y}}(t)^T \mathbf{M} d\mathbf{y} = \frac{1}{2}\dot{\mathbf{y}}(t_k)^T \mathbf{M}\dot{\mathbf{y}}(t_k) - \frac{1}{2}\dot{\mathbf{y}}(0)^T \mathbf{M}\dot{\mathbf{y}}(0) = \frac{1}{2}\dot{\mathbf{y}}_k^T \mathbf{M}\dot{\mathbf{y}}_k \tag{15}$$

where $\dot{\mathbf{y}}(t)$ is the $n \times 1$ absolute velocity vector, $\dot{\mathbf{y}}_k$ represents the discretized form of $\dot{\mathbf{y}}(t_k)$, and t_k represents the k^{th} time step at which the energy value is calculated. The structure is assumed to be at rest when the earthquake begins, and therefore $\dot{\mathbf{y}}(0) = \mathbf{0}$. Due to the squaring of the absolute velocity vector in Eq. (15) and a positive definite **M** matrix, the kinetic energy is always positive.

4.2 Damping Energy (*DE*)

The second integral on the left-hand side of Eq. (14) represents the damping energy (*DE*), which is the energy dissipated via viscous damping mechanism within the structure. The integrand is always positive, and therefore damping energy always accumulates over time. In terms of numerical simulation, the integral can be numerically approximated by evaluating the area underneath the curve using the trapezoidal rule:

$$DE(t_k) = \int_0^{t_k} \dot{\mathbf{x}}(t)^T \mathbf{C} d\mathbf{x} = \sum_{k=1}^{t_k} \frac{1}{2}\left(\dot{\mathbf{x}}_{k-1}^T + \dot{\mathbf{x}}_k^T\right) \mathbf{C} \left(\mathbf{x}_k - \mathbf{x}_{k-1}\right) \tag{16}$$

where $\mathbf{x}_k$ represents the discretized form of $\mathbf{x}(t_k)$ and $\dot{\mathbf{x}}_k$ represents the discretized forms of $\dot{\mathbf{x}}(t_k)$.

4.3 Strain Energy (*SE*)

The third integral on the left-hand side of Eq. (14) represents the strain energy (*SE*) of the moment-resisting frame and can be evaluated as:

$$SE(t_k) = \int_0^{t_k} \mathbf{x}'(t)^T \mathbf{K} d\mathbf{x}' = \frac{1}{2}\mathbf{x}'(t_k)^T \mathbf{K}\mathbf{x}'(t_k) - \frac{1}{2}\mathbf{x}'(0)^T \mathbf{K}\mathbf{x}'(0) = \frac{1}{2}\mathbf{x}_k'^T \mathbf{K}\mathbf{x}_k' \tag{17}$$

where $\mathbf{x}_k'$ represents the discretized form of $\mathbf{x}'(t_k)$. The structure is again assumed to be at rest when the earthquake begins, and therefore $\mathbf{x}'(0) = \mathbf{0}$. Due to the squaring of the elastic displacement vector in Eq. (17) and a positive definite **K** matrix, the strain energy is always positive.

4.4 Plastic Energy (*PE*)

The fourth integral on the left-hand side of Eq. (14), which is associated with inelastic displacements, represents the plastic energy (*PE*) dissipated by the permanent deformations of the structure. Rewriting Eqs. (4) and (5) in the forms:

$$\mathbf{K} d\mathbf{x}'' = \mathbf{K}' d\mathbf{\Theta}'' \quad , \quad \mathbf{x}'(t)^T \mathbf{K}' = \mathbf{m}'(t)^T \tag{18}$$

Then substituting Eq. (18) into the fifth integral of Eq. (14) gives

$$PE(t_k) = \int_0^{t_k} \mathbf{x}'(t)^T \mathbf{K} d\mathbf{x}'' = \int_0^{t_k} \mathbf{x}'(t)^T \mathbf{K}' d\mathbf{\Theta}'' = \int_0^{t_k} \mathbf{m}'(t)^T \, d\mathbf{\Theta}'' = \sum_{i=1}^{m} \int_0^{t_k} m_i'(t) d\theta_i'' = \sum_{i=1}^{m} PE_i(t_k) \tag{19}$$

where PE_i represents the plastic energy dissipation at the i^{th} plastic hinge, $i = 1,...,m$. Through this analytical derivation, it is proven in Eq. (19) that the overall plastic energy dissipation $PE(t_k)$ is exactly equal to the sum of plastic energy dissipation in all the plastic hinges $PE_i(t_k)$.

The term PE_i in Eq. (19) can be numerically approximated by evaluating the area underneath the curve using the trapezoidal rule:

$$PE_i = \int_0^{t_k} m_i'(t) d\theta_i'' = \sum_{k=1}^{t_k} \frac{1}{2}\left(m'_{i,k-1} + m'_{i,k}\right)\left(\theta''_{i,k} - \theta''_{i,k-1}\right) \tag{20}$$

where $m'_{i,k}$ and $\theta''_{i,k}$ represent the discretized forms of $m_i'(t_k)$ and $\theta_i''(t_k)$, respectively. Note that PE_i is computed by integrating the product of elastic moment m_i' and the change in plastic rotation $d\theta_i''$. A positive change in plastic rotation is always caused by a positive moment, and a negative change in plastic rotation is always caused by a negative moment. Therefore, PE_i is always positive and accumulates over time.

4.5 Higher-order Energy (*HE*)

The fifth integral on the left-hand side of Eq. (14) represents higher-order energy (*HE*) due to gravity loads on the structure itself. It is of higher-order because the energy comes from the large *P*-Δ effect, where gravity in the vertical direction is producing energy while going through movement in the horizontal direction. This energy is calculated as follows:

$$HE(t_k) = \int_0^{t_k} \mathbf{x}(t)^T \mathbf{K}_a d\mathbf{x} = \frac{1}{2}\mathbf{x}(t_k)^T \mathbf{K}_a \mathbf{x}(t_k) - \frac{1}{2}\mathbf{x}(0)^T \mathbf{K}_a \mathbf{x}(0) = \frac{1}{2}\mathbf{x}_k^T \mathbf{K}_a \mathbf{x}_k \tag{21}$$

where the structure is again assumed to be at rest when the earthquake begins, and therefore $\mathbf{x}(0) = \mathbf{0}$. Due to the squaring of the total displacement vector in Eq. (21) and a negative definite $\mathbf{K}_a$ matrix as presented in Eq. (9), the higher-order energy is always negative and varies with time.

4.6 Input Energy (*IE*)

Finally, the integral on the right-hand side of Eq. (14) represents the absolute input energy (*IE*) due to the earthquake ground motion, and this integral can be numerically approximated by evaluating the area underneath the ground motion curve using the trapezoidal rule:

$$IE(t_k) = \int_0^{t_k} \ddot{\mathbf{y}}(t)^T \mathbf{M} d\mathbf{g} = \sum_{k=1}^{t_k} \frac{1}{2}\left(\ddot{\mathbf{y}}_{k-1}^T + \ddot{\mathbf{y}}_k^T\right) \mathbf{M} \left(\mathbf{g}_k - \mathbf{g}_{k-1}\right) \tag{22}$$

where $\ddot{\mathbf{y}}_k$ represents the discretized form of $\ddot{\mathbf{y}}(t_k)$ and $\mathbf{g}_k$ represents the discretized form of $\mathbf{g}(t_k)$.

In summary, substituting Eqs. (15), (16), (17), (19), (21), and (22) into Eq. (14), the energy balance equation becomes

$$KE + DE + SE + HE + PE = IE \tag{23}$$

5. Energy Balance Verification

To verify the energy is balanced in Eq. (23), consider a one-story one-bay frame as shown in Fig.1a. Assume an elastic modulus of $E = 200$ GPa and a moment of inertia of $I = 4.16 \times 10^8$ mm^4 for both the beam and columns. Also let the gravity load be $P = 5340$ kN acting on each column, which is equivalent to 13.6 % of the critical buckling load (i.e., $P/P_{cr} = 0.136$). Finally, six plastic hinges are included in the model as shown in Fig.1a with labels from #1 to #6. Assume material exhibits elastic-perfectly-plastic behavior with moment capacities $m_c = 3909$ kN·m for the columns and $m_b = 3130$ kN·m for the beam.

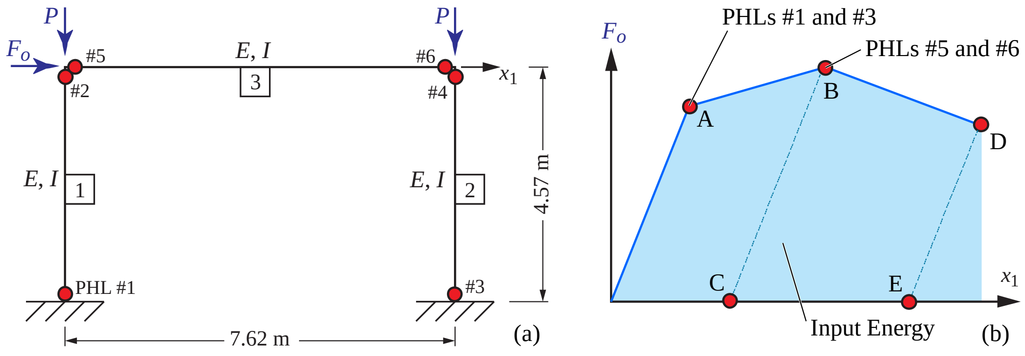


Fig. 1 – One-story one-bay moment resisting frame

No leaning column is used in the model, and therefore $HE = 0$. Now consider a static horizontal load F_o is incrementally applied to the frame. A typical pushover curve for the frame is shown in Fig.1b, where Point A shows the yielding of PHLs #1 and #3, and Point B shows the yielding of PHLs #5 and #6. Input energy (*IE*) is calculated by determining the area underneath the curve as highlighted in blue. With static load being applied, $KE = DE = 0$. Therefore, the energy balance equation in Eq. (23) becomes

$$SE + PE = IE \tag{24}$$

5.1 Loading Up to Point A

First consider Point A shown in Fig.1b, where the moments in PHLs #1 and #3 are exactly equal to the maximum moment capacities. At this point, the applied force is $F_o = 2289$ kN, and the resulting displacement is $x_1 = 0.229$ m as shown in Fig.2a. This gives an elastic stiffness of the frame of $K = F_o / x_1 = 9996$ kN/m. The resultant forces of each member are shown in Fig.2b, and Fig.2c shows that PHLs #1 and #3 are at yield, while PHLs #5 and #6 are below yield. The energy quantities are calculated as:

$$SE = \frac{1}{2}(9996)(0.229)^2 = 262 \text{ kJ} \quad , \quad PE = 0 \text{ kJ} \quad , \quad IE = \frac{1}{2}(0.229)(2289) = 262 \text{ kJ} \tag{25}$$

Therefore, from Eq. (25), $SE + PE = IE$, and Eq. (24) is satisfied.

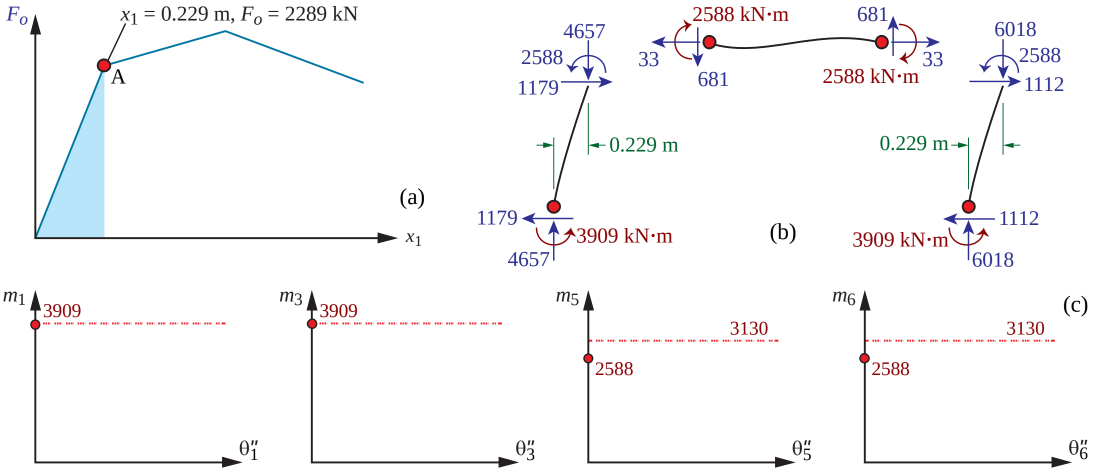


Fig. 2 – State of structure at Point A

5.2 Loading Up to Point B

Now let the applied force be increased up to Point B shown in Fig.1b, where the moments in PHLs #5 and #6 are exactly equal to the moment capacities. At this point, the applied force is $F_o = 2326$ kN, and the resulting displacement is $x_1 = 0.316$ m as shown in Fig.3a. The resultant forces of each member are shown in Fig.3b, and Fig.3c shows that PHLs #5 and #6 are at yield, while PHLs #1 and #3 are beyond yield with plastic rotations of $\theta_1'' = \theta_3'' = 0.0245$ rad. Note that the moments m_1 and m_3 are below the moment capacity of the columns due to axial force and yield moment interactions. Corresponding to the plastic rotations in PHLs #1 and #3, the elastic moments are calculated to be $m_1' = 3923$ kN·m and $m_3' = 2924$ kN·m, and therefore the inelastic moments are $m_1'' = -40$ kN·m and $m_3'' = -62$ kN·m. The energy quantities are calculated as:

$$PE_1 = \frac{3883 + 3923}{2}(0.0245) = 96 \text{ kJ} \quad , \quad PE_3 = \frac{3862 + 3924}{2}(0.0245) = 95 \text{ kJ} \tag{26a}$$

$$SE = \frac{1}{2}\frac{(2326)^2}{9996} = 271 \text{ kJ} \quad , \quad IE = 262 + \frac{(2326 + 2289)}{2}(0.316 - 0.229) = 462 \text{ kJ} \tag{26b}$$

$$PE = PE_1 + PE_3 = 191 \text{ kJ} \tag{26c}$$

Therefore, from Eqs. (26b) and (26c), $SE + PE = IE$, and Eq. (24) is satisfied.

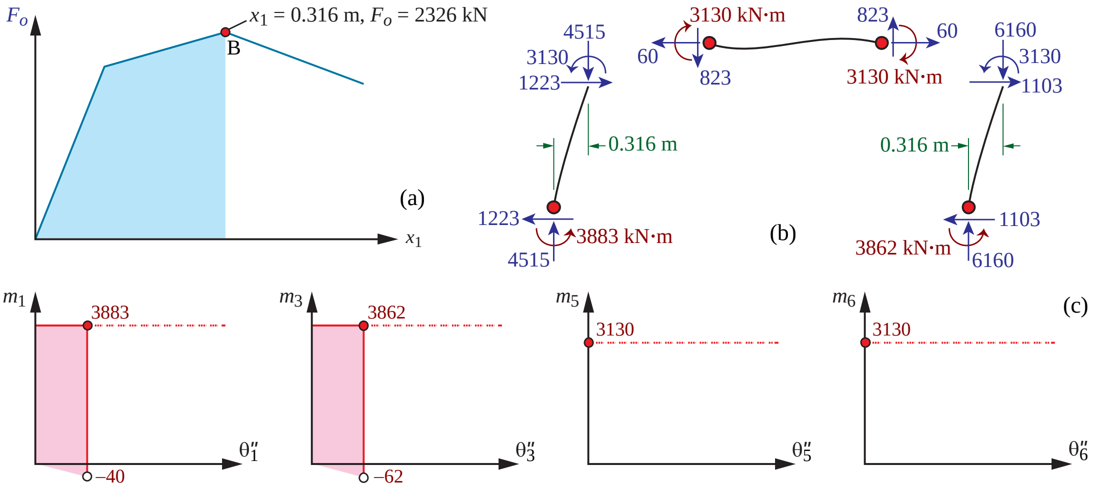


Fig. 3 – State of structure at Point B

5.3 Unloading to Point C

Now let the applied force be completely removed to reach Point C shown in Fig.1b. At this point, the applied force is $F_o = 0$ kN, and the resulting displacement is $x_1 = x_1'' = 0.084$ m (i.e., the elastic displacement $x_1' = 0$) as shown in Fig.4a. The resultant forces of each member are shown in Fig.4b, and Fig.4c shows that PHLs #5 and #6 are elastic, while PHLs #1 and #3 unload to negative moments of $m_1 = m_1'' = -40$ kN·m and $m_3 = m_3'' = -62$ kN·m (i.e., the elastic moments $m_1' = m_3' = -40$). Since plastic energy remains the same during unloading, i.e., $PE_1 = 96$ kJ and $PE_3 = 95$ kJ, the energy quantities are calculated as:

$$SE = 0 \text{ kJ} \quad , \quad PE = PE_1 + PE_3 = 191 \text{ kJ} \quad , \quad IE = 462 - \frac{1}{2}\frac{(2326)^2}{9996} = 191 \text{ kJ} \tag{27}$$

Therefore, from Eq. (27), $SE + PE = IE$, and Eq. (24) is satisfied.

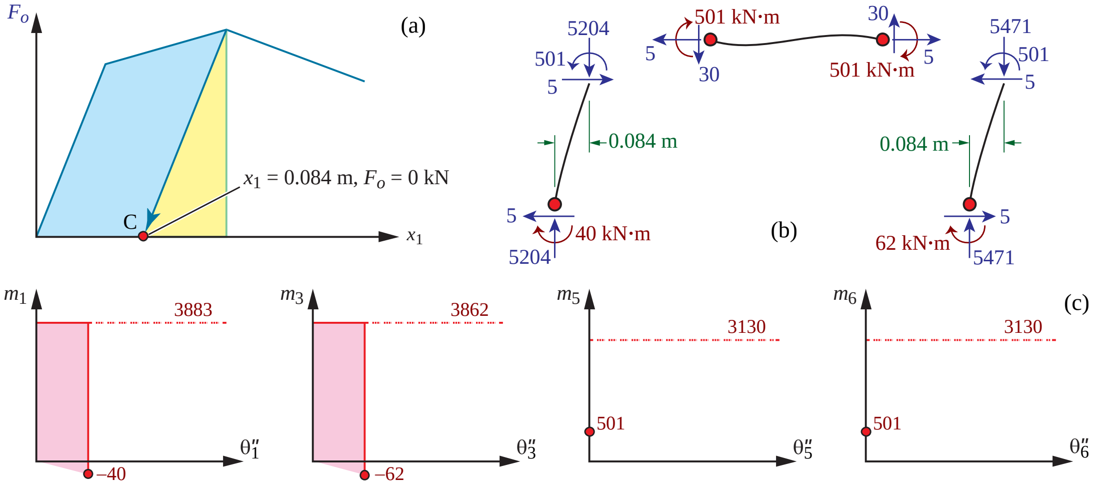


Fig. 4 – State of structure at Point C

5.4 Reloading Up to Point D

Now let the applied force resumes and continues beyond Point B until reaching Point D shown in Fig.1b. At this point, a yield mechanism has formed, which means any additional load can cause instability of the frame. For this reason, displacement control is used. Let the displacement of the frame be $x_1 = 0.457$ m, and the corresponding applied force is calculated to be $F_o = 1995$ kN as shown in Fig.5a. The resultant forces of each member are shown in Fig.5b, and Fig.5c shows that PHLs #1, #3, #5, and #6 are beyond yield with plastic rotations of $\theta_1'' = \theta_3'' = 0.0556$ rad and $\theta_5'' = \theta_6'' = 0.0309$ rad. Corresponding to these plastic rotations, the inelastic moments are shown in Fig.5c. The energy quantities are then calculated as:

$$PE_1 = 96 + (3883 + 40 + 3883 - 515)(0.0556 - 0.0245)/2 = 209 \text{ kJ} \tag{28a}$$

$$PE_3 = 95 + (3862 + 62 + 3862 - 493)(0.0556 - 0.0245)/2 = 208 \text{ kJ} \tag{28b}$$

$$PE_5 = PE_6 = (3130 - 501 + 3130 - 873)(0.0309)/2 = 75.5 \text{ kJ} \tag{28c}$$

$$PE = PE_1 + PE_3 + PE_5 + PE_6 = 568 \text{ kJ} \tag{28d}$$

$$SE = \frac{1}{2}\frac{(1995)^2}{9996} = 199 \text{ kJ} \quad , \quad IE = 462 + \frac{1}{2}(1995 + 2326)(0.457 - 0.316) = 767 \text{ kJ} \tag{28e}$$

Therefore, from Eqs. (28d) and (28e), $SE + PE = IE$, and Eq. (24) is satisfied.

5.5 Unloading to Point E

Finally, let the applied force be completely removed to reach Point E shown in Fig.1b. At this point, the applied force is $F_o = 0$ kN, and the resulting displacement is $x_1 = x_1'' = 0.258$ m as shown in Fig.5a. The resultant forces of each member are shown in Fig.5b, and Fig.5c shows the moment-rotation plot of each plastic hinge. Since the strain energy at zero force is $SE = 0$ kJ, and the plastic energy remains the same during unloading, i.e., $PE_1 = 209$ kJ, $PE_3 = 208$ kJ, and $PE_5 = PE_6 = 75.5$ kJ, the energy quantities are calculated as:

$$SE = 0 \text{ kJ} \quad , \quad PE = PE_1 + PE_3 + PE_5 + PE_6 = 568 \text{ kJ} \quad , \quad IE = 767 - \frac{1}{2}\frac{(1995)^2}{9996} = 568 \text{ kJ} \tag{29}$$

Therefore, from Eq. (29), $SE + PE = IE$, and Eq. (24) is satisfied.

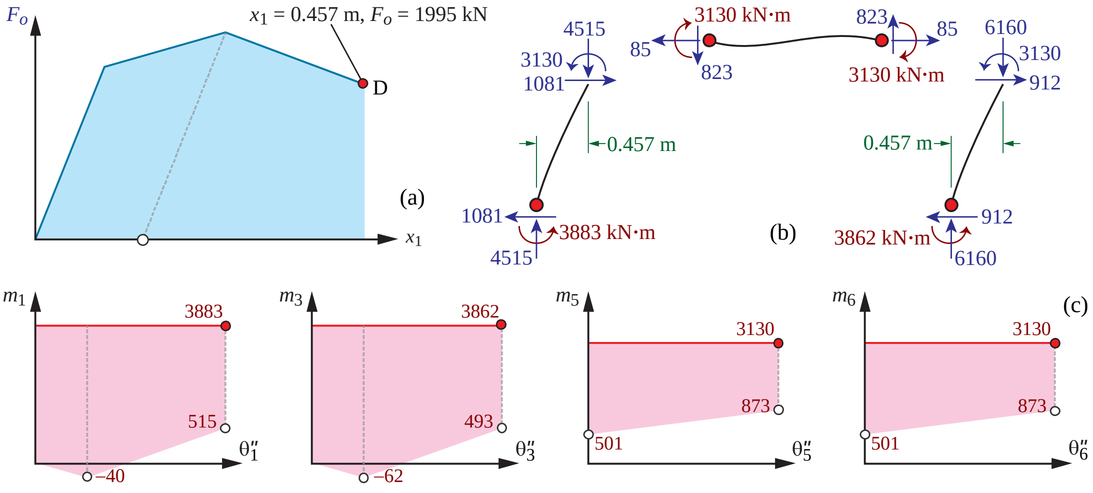


Fig. 5 – State of structure at Point D

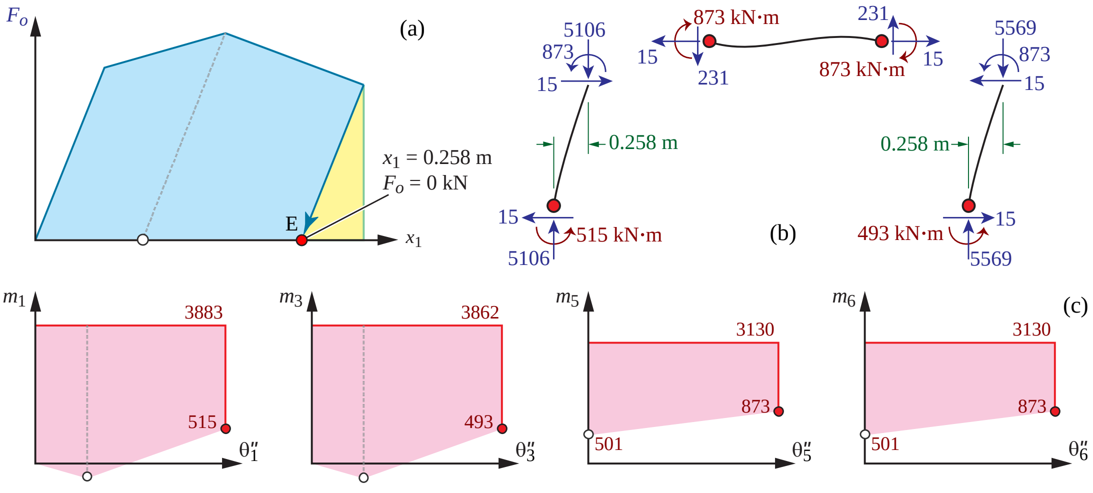


Fig. 6 – State of structure at Point E

6. Numerical Simulation of a 4-story Moment-Resisting Frame

To demonstrate the balance of energy in dynamic analysis, consider the four-story moment-resisting frame as shown in Fig.7a designed according to code [8]. This frame contains 36 DOFs (i.e., $n = 36$), of which 4 are horizontal translations, 16 are vertical translations, and 16 are joint rotations. Also, this frame has a total of 56 PHLs (i.e., $m = 56$). Assume each floor has a mass of 95 000 kg, and gravity load of 431 kN is applied on each exterior column member and 632 kN is applied on each interior column member as shown in Fig.7b. Based on these gravity load on the frame, the stiffness matrices $\mathbf{K}$, $\mathbf{K}'$, and $\mathbf{K}''$ in Eq. (4) are constructed such that geometric nonlinearity is included in the formulation. In addition, a leaning column is used to account for all the gravity loads from other parts of the structure. Let the gravity loads on the leaning column be 3026 kN per floor. Finally, a 2 % damping is assumed in all four modes of vibration.

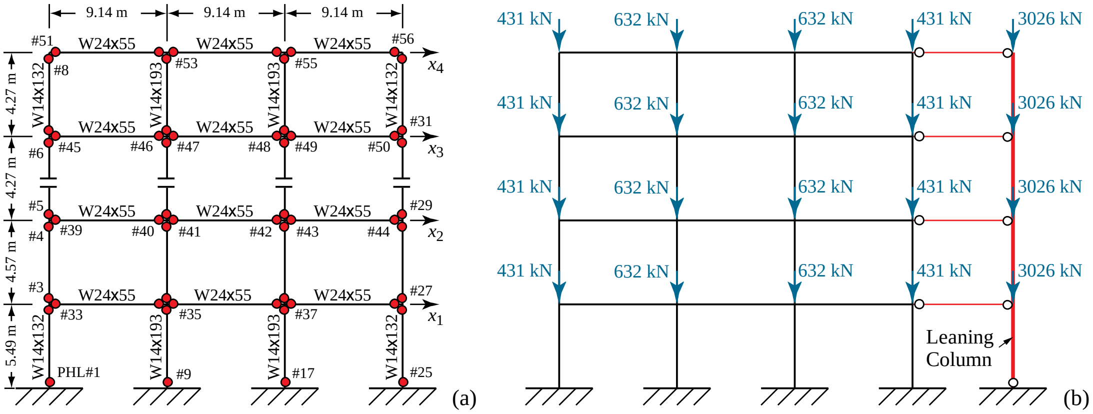


Fig. 7 – Four-story moment-resisting steel frame and corresponding gravity loads

Assume the yield stress of the member is 345 MPa and all 56 plastic hinges exhibit elastic-perfectly-plastic behavior. By subjecting the frame to the 1995 Kobe earthquake ground acceleration shown in Fig.8, the energy response histories are summarized in Fig.9. The results confirm that *KE* and *SE* are always positive (see Fig.9a), *DE* and *PE* are cumulative (see Fig.9b), and *HE* is always negative (see Fig.9d). Of particular interest is Fig.9c, which shows the plot of *IE* and *PE*. This figure shows that significant portion of input energy is dissipated through plastic energy, indicating that significant damage occurs in the structure.

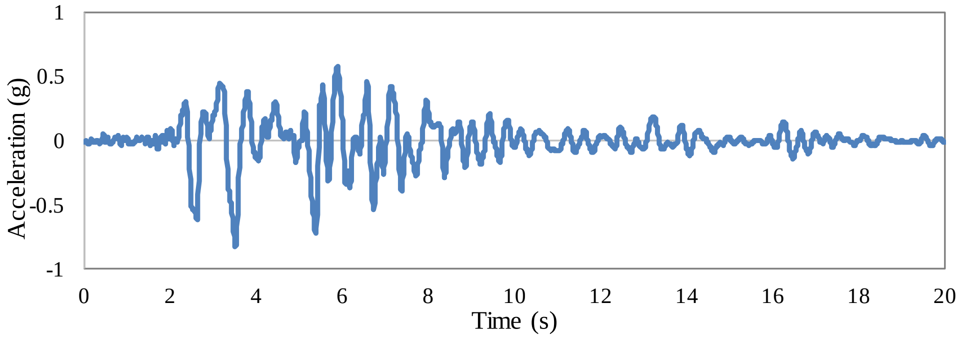


Fig. 8 – Recorded 1995 Kobe earthquake ground motion at Kajima station Component 000

In terms of plastic energy dissipation at individual plastic hinges, Fig.10a shows the maximum plastic energy at each plastic hinge, i.e., PE_i . Since plastic energy accumulates over time, the maximum plastic energy always occurs at the end of the earthquake duration. In addition, summing the plastic energy at all the plastic hinges in Fig.10a gives 950 kJ, which is equal to the total plastic energy dissipation *PE* at the end of the earthquake shown in Fig.9c.

Now consider the same 4-story frame as shown in Fig.7 but is subjected to the Kobe earthquake shown in Fig.8 with a scale factor of 1.4. The energy response histories are shown in Fig.11, where Fig.11a shows the storing energy quantities including *KE*, *SE*, and *HE*, while Fig.11b shows the cumulative energy quantities including *DE*, *PE*, and *IE*. Comparison of results in Fig.11 with those in Fig.9 shows that both *PE* and *IE* increase due to a larger earthquake while other energy quantities remain practically the same. This is consistent with intuition that larger input energy due to larger earthquake must be dissipated through plastic energy, resulting in more structural damage. The plastic energy dissipated in each plastic hinge, i.e., PE_i , is shown in Fig.10b. Summing these individual plastic energy values in Fig.10b gives 1484 kJ, which is equal to the total plastic energy dissipation *PE* at the end of the earthquake shown in Fig.11b.

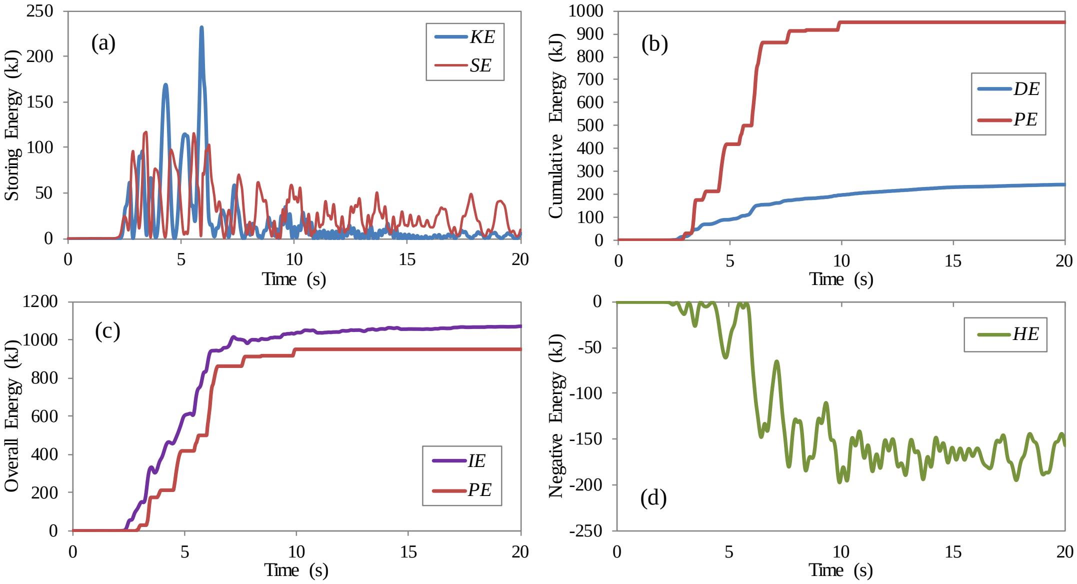


Fig. 9 – Energy responses of the 4-story framed structure due to Kobe earthquake

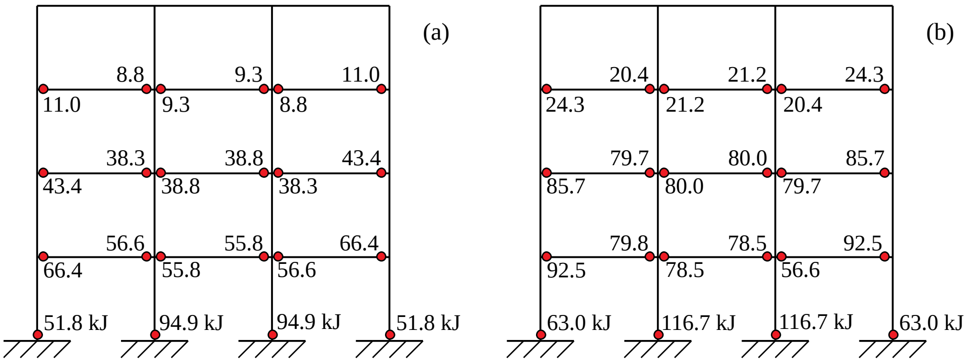


Fig. 10 –Plastic energy dissipation among individual plastic hinges

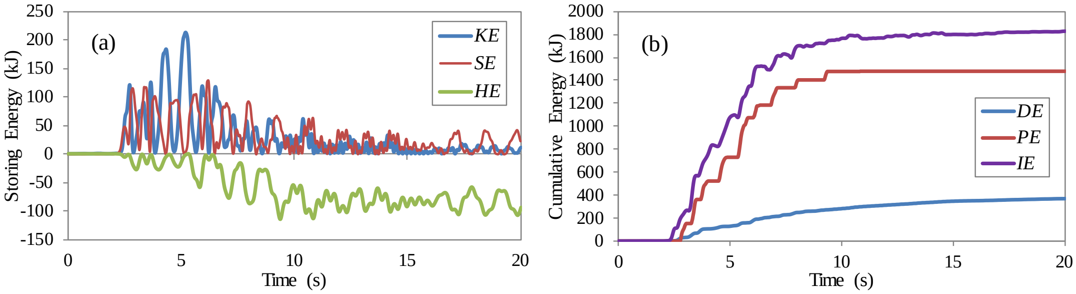


Fig. 11 – Energy responses of the 4-story framed structure due to 1.4 × Kobe earthquake

7. Conclusion

The analytical theory of seismic energy balance and the associated computational method for evaluating the seismic energy in structures are presented with the inclusion of geometric nonlinearity and material nonlinearity. This method successfully separates the coupling effect of material nonlinearity and geometric nonlinearity by using inelastic displacement. By expressing the input energy as the sum of kinetic energy, damping energy, strain energy, higher-order energy, and plastic energy, the energy representation of the structural response due to earthquake ground motion is complete. In particular, the two nonlinear energies are: (1) Higher-order energy (HE), which represents the negative energy stored in the structure that can be detrimental to structural stability; and (2) Plastic energy (PE), which represents the dissipation of energy and reduction of structural response due to material nonlinearity.

Due to static loading, the summation of these energy forms is exactly equal to the input energy. This is verified step by step in the present paper. The key is that plastic energy requires to use of elastic moment in the calculation, which is contrary to many past research [9-13] that uses total moment to calculate the area underneath the moment-rotation plots. The study is then extended to dynamic analysis, where plastic energy dissipation among individual plastic hinges is calculated. By using the elastic moment in the calculation of plastic energy, it is shown that energy is balanced in a dynamic system. This balance of energy is important because it gives a higher level of confidence in quantifying the energy demand in each component (i.e., PE_i), which in turn can be compared with the energy capacity of each component in energy-based design.

8. References

[1] Tembulkar J, Nau JM (1987): Inelastic modeling and seismic energy dissipation. *Journal of Structural Engineering ASCE*, **113** (6), 1373-1377.

[2] Fajfar P, Vidic T, Fischinger M (1989): Seismic demand in medium-period and long-period structures. *Earthquake Engineering and Structural Dynamics*, **18** (8), 1133-1144.

[3] Fajfar P (1992): Equivalent ductility factors, taking into account low-cycle fatigue. *Earthquake Engineering and Structural Dynamics*, **21** (10), 837-848.

[4] Chai YH (2005): Incorporating low-cycle fatigue model into duration-dependent inelastic design spectra. *Earthquake Engineering and Structural Dynamics*, **34** (1), 83-96.

[5] Uang CM, Bertero VV (1990): Evaluation of seismic energy in structures. *Earthquake Engineering and Structural Dynamics*, **19** (1), 77-90.

[6] Wong KKF, Yang R (2002): Earthquake response and energy evaluation of inelastic structures. *Journal of Engineering Mechanics ASCE*, **128** (3), 308-317.

[7] Wong KKF, Yang R (1999): Inelastic dynamic response of structures using force analogy method. *Journal of Engineering Mechanics ASCE*, **125** (10), 1190-1199.

[8] Harris JL, Speicher MS (2015): *Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings Volume 1: Special Moment Frames*, National Institute of Standards and Technology, USA.

[9] Park YJ, Ang AHS (1985): Mechanistic seismic damage model for reinforced concrete. *Journal of Structural Engineering ASCE*, **111** (4), 722-739.

[10] McCabe SL, Hall WJ (1989): Assessment of seismic structural damage. *Journal of Structural Engineering ASCE*, **115** (9), 2166-2183.

[11] Chai YH, Fajfar P (2000): A procedure for estimating input energy spectra for seismic design. *Journal of Earthquake Engineering*, **4** (4), 539-561.

[12] Bojorquez E, Reyes-Salazar A, Teran-Gilmore A, Ruiz SE (2010): Energy-based damage index for steel structures. *Steel and Composite Structures*, **10** (4), 331-348.

[13] Benavent-Climent A (2011): An energy-based method for seismic retrofit of existing frames using hysteretic dampers. *Soil Dynamics and Earthquake Engineering*, **31** (10), 1385-1396.

The 2019 NIST Speaker Recognition Evaluation CTS Challenge

Seyed Omid Sadjadi[1], Craig Greenberg[1], Elliot Singer[2,†], Douglas Reynolds[2,†], Lisa Mason[3], Jaime Hernandez-Cordero[3]

[1]NIST ITL/IAD/Multimodal Information Group, MD, USA
[2]MIT Lincoln Laboratory, MA, USA
[3]U.S. Department of Defense, MD, USA

craig.greenberg@nist.gov

Abstract

In 2019, the U.S. National Institute of Standards and Technology (NIST) conducted a leaderboard style speaker recognition challenge using conversational telephone speech (CTS) data extracted from the unexposed portion of the Call My Net 2 (CMN2) corpus previously used in the 2018 Speaker Recognition Evaluation (SRE). The SRE19 CTS Challenge was organized in a similar manner to SRE18, except it offered only the *open* training condition. In addition, similar to the NIST i-vector challenge, the evaluation set consisted of two subsets: a *progress* subset, and a *test* subset. Trials for the *progress* subset comprised 30% of the target speakers from the unexposed portion of the CMN2 corpus and was used to monitor progress on the leaderboard, while trials from the remaining 70% of the speakers were allocated for the *test* subset, which was used to generate the official final results determined at the end of the challenge. Which subset (i.e., *progress* or *test*) a trial belonged to was unknown to challenge participants, and each system submission had to contain outputs for all of the trials. The SRE19 CTS Challenge also served as a prerequisite for entrance to the main SRE19 whose primary task was audio-visual person recognition. A total of 67 organizations (forming 51 teams) from academia and industry participated in the CTS Challenge and submitted 1347 valid system outputs. This paper presents an overview of the evaluation and several analyses of system performance for all primary conditions in the CTS Challenge. Compared to the CTS track of SRE18, the SRE19 CTS Challenge results indicate remarkable improvements in performance which are mainly attributed to 1) the availability of large amounts of in-domain development data (publicly available and/or proprietary) from a large number of labeled speakers, 2) speaker representations (aka embeddings) extracted using extended and more complex end-to-end neural network frameworks, and 3) effective use of the provided large development set.

1. Introduction

The United States National Institute of Standards and Technology (NIST) organized the 2019 Speaker Recognition Evaluation (SRE19) in the summer–fall of 2019. It was the latest in the ongoing series of speaker recognition technology evaluations conducted by NIST since 1996 [1, 2]. The objectives of the evaluation series are 1) for NIST to effectively measure system-calibrated performance of the current state of technology, 2) to provide a common test bed that enables the research to explore promising new ideas in speaker recognition, and 3) to support the community in their development of advanced technology incorporating these ideas. The basic task in the NIST SREs is speaker detection, that is, determining whether a specified target speaker is talking in a given test speech recording.

SRE19 consisted of two separate activities: 1) a leaderboard-style challenge using conversational telephone speech (CTS) extracted from the unexposed portions of the Call My Net 2 (CMN2) corpus collected by the Linguistic Data Consortium (LDC), which was also previously used to extract the SRE18 CTS development and test sets, and 2) a regular evaluation using audio-visual material extracted from the unexposed portions of the Video Annotation for Speech Technology (VAST) corpus [3], also collected by the LDC. This paper describes the task, the performance metric, data, and the evaluation protocol as well as results and performance analyses of submissions for the SRE19 CTS Challenge. The Audio-Visual SRE19 overview and results is described in another paper [4]. It is worth noting here that the CTS challenge also served as a prerequisite for the audio-visual evaluation, meaning that in order to participate in the regular evaluation, one must have first completed the challenge (i.e., submitted to NIST valid system outputs along with sufficiently detailed system description reports). SRE19 was coordinated entirely online using a freshly designed web platform[1] deployed on Amazon Web Services (AWS)[2] that supported a variety of evaluation-related services such as registration, data license agreement management, data distribution, system output submission and validation/scoring, and system description/presentation uploads.

The SRE19 CTS Challenge was organized in a similar manner to the CTS track of SRE18 [5], except it only offered the *open* training condition in which participants were allowed to use any publicly available and/or proprietary data for system training and development purposes. In addition, a much larger development set was released for the SRE19 CTS Challenge which contained the entire SRE18 development and test sets including segments from 213 labeled speakers as well as segments from more than 1000 unlabeled speakers. Furthermore, similar to the NIST i-vector speaker recognition challenge [6], the evaluation set consisted of two subsets: a *progress* subset, and a *test* subset. Trials for the *progress* subset comprised 30% of the target speakers from the unexposed portion of the CMN2 corpus

[†] The work of MIT Lincoln Laboratory is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1] https://sre.nist.gov

[2] see Disclaimer.

Sadjadi, Omid; Greenberg, Craig; Singer, Elliot; Reynolds, Douglas; Mason, Lisa; Hernandez-Cordero, Jaime. "The 2019 NIST Speaker Recognition Evaluation CTS Challenge." Paper presented at The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, JP. November 01, 2020 - November 05, 2020.

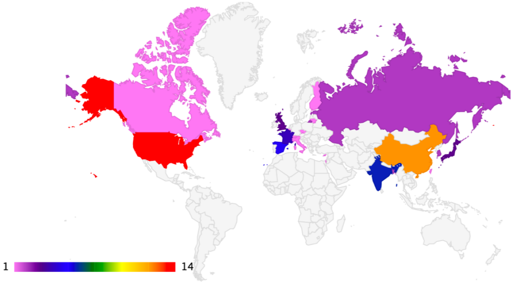


Figure 1: *Heat map of the world countries showing the number of SRE19 CTS Challenge participating sites per country.*

and was used to monitor progress on the leaderboard, while trials from the remaining 70% of the speakers were allocated for the *test* subset, which was used to generate the official final results determined at the end of the challenge. Which subset (i.e., *progress* or *test*) a trial belonged to was unknown to challenge participants, and each system submission had to contain outputs for all of the trials. The participants could make multiple submissions (up to 3 per day), and the leaderboard displayed the best submission performance results thus far received and processed. Over the course of the challenge, which ran from July 15 through October 7, 2019, a total of 51 teams, 23 of which were led by industrial institutions, from 67 sites made 1347 valid submissions (note that the participants processed the data locally and submitted only the output of their systems to NIST for scoring and analysis purposes). Figure 1 displays a heatmap representing the number of participating sites per country. It should be noted that all participant information, including country, was self-reported. The number of submissions per team in the SRE19 CTS Challenge is shown in Figure 2.

Finally, as in SRE18, and in an effort to provide a reproducible state-of-the-art baseline for the SRE19 CTS Challenge, NIST released well in advance of the evaluation period a report [7] containing the baseline speaker recognition system description and results obtained using a state-of-the-art (as of SRE18) deep neural network (DNN) embedding based system (see Section 5 for more details).

2. Task Description

The task for the SRE19 CTS Challenge was *speaker detection*, meaning given a segment of speech and the target speaker enrollment data, automatically determine whether the target speaker is speaking in the segment. A segment of speech (test segment) along with the enrollment speech segment(s) from a designated target speaker constitute a *trial*. The system is required to process each trial independently and to output a log-likelihood ratio (LLR), using natural (base e) logarithm, for that trial. The LLR for a given trial including a test segment s is defined as follows

$$LLR(s) = \log\left(\frac{P\left(s|H_0\right)}{P\left(s|H_1\right)}\right), \tag{1}$$

where $P\left(\cdot\right)$ denotes the probability distribution function (pdf), and H_0 and H_1 represent the null (i.e., s is spoken by the enrollment speaker) and alternative (i.e., s is not spoken by the enrollment speaker) hypotheses, respectively.

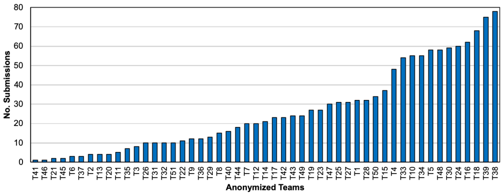


Figure 2: *Submission statistics for the SRE19 CTS Challenge.*

3. Data

In this section we provide a brief description of the data released in the SRE19 CTS Challenge for system training, development, and test.

3.1. Training set

As noted previously, unlike in SRE18 which offered *fixed* and *open* training conditions, the SRE19 CTS Challenge only offered the *open* training condition that allowed the use of any publicly available and/or proprietary data for system training and development purposes. The motivation behind this decision was twofold. First, results from the most recent NIST SREs (i.e., SRE16 [8] and SRE18) indicated limited performance improvements, if any, from unconstrained training compared to *fixed* training, although participants had cited lack of time and/or resources during the evaluation period for not demonstrating significant improvement with *open* versus *fixed* training. Second, the number of publicly available large-scale data resources for speaker recognition has dramatically increased over the past few years (e.g., see VoxCeleb[3]). Therefore, removing the *fixed* training condition would allow more in-depth exploration into the gains that could be achieved with the availability of unconstrained resources given the success of data-hungry Neural Network based approaches in the most recent evaluation (i.e. SRE18 [5]). Nevertheless, it is worth noting here that during the discussion sessions at the post-evaluation workshop, which was held in December 2019 in Singapore, several participating teams requested the re-introduction of the *fixed* training condition to facilitate meaningful and fair cross-system comparisons in terms of core speaker recognition algorithms/approaches (as opposed to particular data) used to develop the systems.

Although SRE19 allowed unconstrained system training and development, participating teams were required to provide a sufficient description of speech and non-speech (e.g., noise samples, room impulse responses, and filters) data resources as well as pre-trained models used during the training and development of their systems.

3.2. Development and evaluation sets

For the sake of convenience, in particular for new SRE participants, NIST provided an *in-domain* development set that could be used for both system training and development purposes. This Development set simply combined the SRE18 CTS development and test sets into one package (i.e. LDC2019E59). Participants could obtain this dataset through the evaluation web platform (`https://sre.nist.gov`) after signing the LDC data license agreement. The first three rows in Table 1 summarize the statistics for this development set.

[3] `http://www.robots.ox.ac.uk/~vgg/data/voxceleb/`

Table 1: *Statistics for the SRE19 CTS Challenge development (DEV) and evaluation (EVAL), i.e., progress and test sets*

Set	Dev/Test	#speakers (M / F)	#1-segment enrollment	#3-segment enrollment	#Test segments	#target/non-target trials
CTS'18 (DEV)	Dev-labeled	9 / 16	100	25	1566	7830 / 100,265
	Dev-unlabeled	–	–	–	2332	–
	Test	70 / 118	752	188	12,135	19,298 / 2,002,332
CTS'19 (EVAL)	Progress	21 / 37	232	58	4066	20,330 / 618,360
	Test	49 / 88	547	137	9515	47,518 / 2,000,000

Table 2: *Primary partitions in the CTS Challenge* ***progress*** *set*

Partition	Elements	#target	#non-target
Gender	male	7095	141,900
	female	13,235	476,460
#enrollment segments	1	16,264	494,688
	3	4066	123,672
Phone# match	Y	9452	0
	N	10,878	618,320
CTS type	PSTN	15,935	484,700
	VoIP	4395	133,660

Table 3: *Primary partitions in the CTS Challenge* ***test*** *set*

Partition	Elements	#target	#non-target
Gender	male	15,843	433,078
	female	31,675	1,569,090
#enrollment segments	1	38,003	1,600,661
	3	9515	401,507
Phone# match	Y	24,456	0
	N	23,062	2,000,000
CTS type	PSTN	36,308	1,536,768
	VoIP	11,210	465,400

The speech segments in the SRE19 CTS Challenge development (*DEV*) and evaluation (*EVAL*) sets were extracted from the CMN2 corpus collected by the LDC to support speech technology evaluations. The CMN2 corpus consists of CTS recordings spoken in Tunisian Arabic, which were collected over the traditional Public Switched Telephone Network (PSTN) and the more recent Voice over IP (VOIP) platforms outside North America. For CMN2 data collection, the LDC recruited a few hundred speakers called *claques* who made multiple calls to people in their social network (e.g., family, friends). Claques were encouraged to use different telephone instruments (e.g., cell phone, landline) in a variety of settings (e.g., noisy cafe, quiet office) for their initiated calls and were instructed to talk for at least 8–10 minutes on a topic of their choice. All CMN2 recordings are encoded as a-law sampled at 8 kHz in SPHERE [9] formatted files.

Similar to the most recent SREs (i.e., SRE16 and SRE18), there were two enrollment scenarios for the SRE19 CTS Challenge, namely 1-segment and 3-segment conditions. As the names imply, in the 1-segment condition only one approximately 60 s speech segment was given for enrollment, while in the 3-segment condition three approximately 60 s speech segments (from the same phone number) were provided to build the model of the target speaker. It is worth noting that the 3-segment condition only involved the PSTN data, because the number of VoIP calls per *claque* was limited. As part of the *dev* set for the SRE19 CTS Challenge, an *unlabeled* set of 2332 segments (with speech duration uniformly distributed in 10 s to 60 s range) was also made available by NIST. The *unlabeled* segments were extracted from the non-*claque* side of the PSTN/VoIP calls.

For the SRE19 CTS Challenge, the evaluation trials were divided into two subsets: a *progress* subset, and a *test* subset. Trials for the *progress* subset comprised 30% of the target speakers from the unexposed portion of the CMN2 corpus and was used to monitor progress on the leaderboard, while trials from the remaining 70% of the speakers were allocated for the *test* subset which was used to generate the official final results determined at the end of the challenge. The challenge test conditions were as follows:

- The speech durations of the test segments were uniformly sampled ranging approximately from 10 seconds to 60 seconds.
- Trials were conducted with test segments from both same and different phone numbers as the enrollment segment(s).
- There were no cross-gender trials.

The last two rows of Table 1 show the statistics for the SRE19 CTS Challenge *progress* and *test* subsets.

4. Performance Measurement

Similar to the past SREs, the primary performance measure for the SRE19 CTS Challenge was a detection cost defined as a weighted sum of false-reject (miss) and false-accept (false-alarm) error probabilities. Equation (2) specifies the CTS Challenge primary normalized cost function for some decision threshold θ,

$$C_{norm}\left(\theta\right) = P_{miss}\left(\theta\right) + \beta \times P_{fa}\left(\theta\right), \qquad (2)$$

where β is defined as

$$\beta = \frac{C_{fa}}{C_{miss}} \times \frac{1 - P_{target}}{P_{target}}. \qquad (3)$$

The parameters C_{miss} and C_{fa} are the cost of a missed detection and cost of a false-alarm, respectively, and P_{target} is the *a priori* probability that the test segment speaker is the specified target speaker. The primary cost metric, $C_{primary}$ for the CTS Challenge was the average of normalized costs calculated at two points along the detection error trade-off (DET) curve [10], with $C_{miss} = C_{fa} = 1$, $P_{target} = 0.01$ and $P_{target} = 0.005$. Here, $\log(\beta)$ was applied as the detection

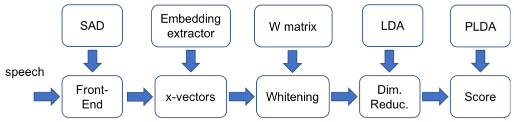


Figure 3: *A simplified block diagram of the baseline speaker recognition system for the SRE19 CTS Challenge.*

threshold θ for computing the actual detection costs. Additional details can be found in the SRE19 CTS Challenge evaluation plan [11].

Similar to the recent SREs (i.e., SRE16 and SRE18), the test data was divided into 16 partitions. Each partition is defined as a combination of: speaker gender (male vs female), number of enrollment segments (1 vs 3), enrollment-test phone number match (Yes vs No), and CTS source type (PSTN vs VoIP). However, because no actual "phone number" metadata was available for either enrollment or test segments extracted from the VoIP calls, the phone number match field only contained "N" for those calls, thereby reducing the effective number of partitions to 12. Also, all non-target trials are from the different (as opposed to the same) phone number partition, assuming each phone number would be only used by one individual. More information about the various partitions in the SRE19 CTS Challenge *progress* and *test* subsets can be found in Tables 2 and 3. $C_{primary}$ was calculated for each partition, and the final result was the average of all the partitions' $C_{primary}$'s.

Also, a minimum detection cost was computed by using the detection thresholds that minimized the detection cost. Note that for minimum cost calculations, the counts for each condition set was equalized before pooling and cost calculation, that is, the minimum cost was computed using a single threshold not one per condition set.

5. Baseline system

In this section, we describe the x-vector baseline system setup including speech and non-speech data used for training the system components as well as the hyper-parameter configurations used in our evaluations. Figure 3 shows a block diagram of the x-vector baseline system. The x-vector system is built using Kaldi [12] (for x-vector extractor training) and the NIST SLRE toolkit for back-end scoring.

5.1. Data

The x-vector baseline system was developed using the data recipes available at `https://github.com/kaldi-asr/kaldi/tree/master/egs/sre16/v2` as well as `https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2`. The x-vector extractor for the *progress* set was trained entirely using speech data extracted from combined VoxCeleb 1 and 2 corpora, while the x-vector extractor for the *test* set used the prior SRE data (i.e., SRE04-10 as in the Kaldi `sre16` recipe) in addition to the combined VoxCeleb. This was done to ensure the baseline results would serve as a fair comparison point for the first time participants who might only have access to the VoxCeleb data, but not to the prior SRE data. In order to increase the diversity of the acoustic conditions in the training set, a 5-fold augmentation strategy was used that added four corrupted copies of the original recordings to the training list. The recordings were corrupted by either digitally adding noise (i.e., babble, general noise, music) or convolving with simulated and measured room impulse responses (RIR). The noise and RIR samples are freely available from `http://www.openslr.org` (see [13] for more details).

All recordings are downsampled to 8 kHz using `sox`.

5.2. Configuration

For speech parameterization, we extracted 23-dimensional MFCCs (including c0) from 25 ms frames every 10 ms using a 23-channel mel-scale filterbank spanning the frequency range 20 Hz–3700 Hz. Before dropping the non-speech frames using an energy based SAD, a short-time cepstral mean subtraction was applied over a 3-second sliding window.

For x-vector extraction, an extended TDNN with 12 hidden layers and rectified linear unit (RELU) non-linearities was trained to discriminate among the speakers in the training set. After training, embeddings were extracted from the 512-dimensional affine component of the 11th layer (i.e., the first segment-level layer). More details regarding the DNN architecture (e.g., the number of hidden units per layer) and the training process can be found in [14].

Prior to dimensionality reduction through LDA (to 250), 512-dimensional x-vectors were centered, whitened, and unit-length normalized. The centering and whitening statistics were computed using the in-domain development data (i.e., LDC2019E59). For backend scoring, a Gaussian PLDA model with a full-rank Eigenvoice subspace was trained using the x-vectors extracted from either 170 k concatenated speech segments from the combined VoxCeleb sets (for the *progress* set), or 50 k speech segments from prior SRE data (for the *test* set), as well as one corrupted version randomly selected from {babble, noise, music, reverb}. The PLDA parameters were then adapted on the in-domain development data (i.e., LDC2019E59) using Bayesian maximum *a posteriori* (MAP) estimation.

Finally, the PLDA verification scores were post-processed using an adaptive score normalization (AS-Norm) scheme proposed in [15]. We used LDC2019E59 as the cohort set, and selected the top 10% of sorted cohort scores for calculating the normalization statistics.

It is worth emphasizing that the configuration parameters employed to build the baseline system are commonly used by the speaker recognition community, and no attempt was made to tune the hyperparameters or data lists utilized to train the models.

6. Results and Discussion

In this section we present some key results and analyses for SRE19 CTS Challenge submissions, in terms of minimum and actual costs as well as DET performance curves.

Figure 4 shows performance of the best submissions per team per subset as well as performance of the baseline systems [7] in terms of the actual and minimum costs, for the SRE19 CTS Challenge *progress* and *test* subsets, respectively. Baseline 1 and 2 denote the baseline speaker recognition systems trained without and with prior SRE data, respectively (see Section 5 for more details). Here, the y-axis limit is set to 0.5 to facilitate cross-system comparisons in the lower cost region. Several observations can be made from the two plots. First, performance trends on the two subsets are generally similar, although slightly better results are observed on the *progress* subset compared to the *test* subset, a phenomenon which is speculated to primarily result from overtuning/overfitting of the submission systems on the *progress* set. Second, nearly half of

Sadjadi, Omid; Greenberg, Craig; Singer, Elliot; Reynolds, Douglas; Mason, Lisa; Hernandez-Cordero, Jaime. "The 2019 NIST Speaker Recognition Evaluation CTS Challenge." Paper presented at The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, JP. November 01, 2020 - November 05, 2020.

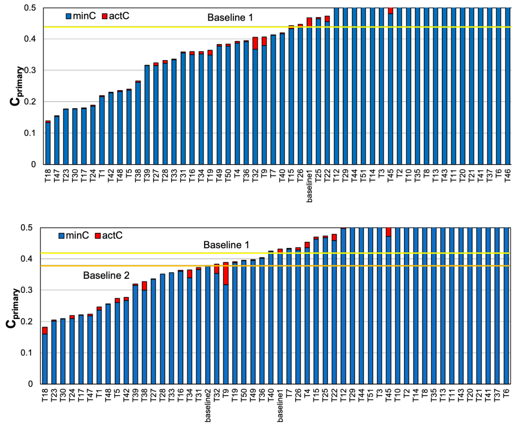


Figure 4: *Performance of the SRE19 CTS Challenge submissions in terms of actual (red) and minimum (blue) costs for the progress (top) and test (bottom) subsets.*

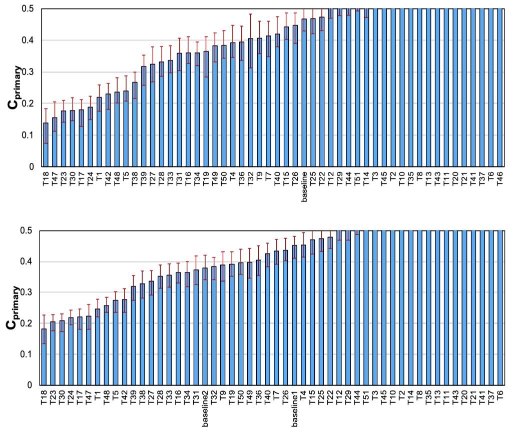


Figure 6: *Performance confidence intervals (95%) of the SRE19 CTS Challenge submissions for the progress (top) and test (bottom) subsets.*

the submissions outperform the baseline system trained on VoxCeleb (i.e., baseline 1), while the number is smaller when compared to the baseline that utilizes the prior SRE data. Third, a majority of the systems achieve relatively small calibration errors, in particular on the *progress* subset. This is in line with the calibration performance of the submitted systems observed for the SRE18 CTS domain. Finally, it can be seen from the figures that, except for the top performing team, the performance gap among the next top-5 teams is not remarkable. A statistical analysis of performance (e.g., confidence intervals for the cost estimates) that sheds more light on actual performance differences among the top performing systems follows later in this section.

Compared to the most recent SRE (i.e., SRE18), there is a notable improvement in speaker recognition performance. Figure 5 presents a performance comparison of SRE18 versus SRE19 CTS submissions for several top performing systems, in terms of actual and minimum detection costs. Performance improvements as large as 70% are achieved by some leading systems, while for others more moderate, but consistent, improvements are observed. These performance improvements are largely attributed to 1) the availability of large amounts of in-domain development data from a large number of labeled speakers (e.g., the entire SRE18 CTS development and test data, or other proprietary in-domain data), and 2) the use of extended and more complex end-to-end neural network frameworks for speaker embedding extraction that can effectively exploit vast amounts of training data made available through data augmentation and/or large-scale datasets such as VoxCeleb[3].

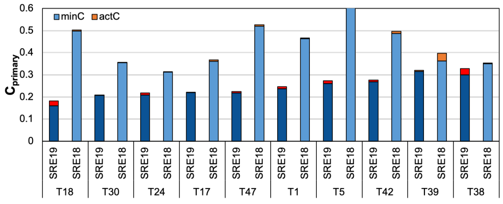
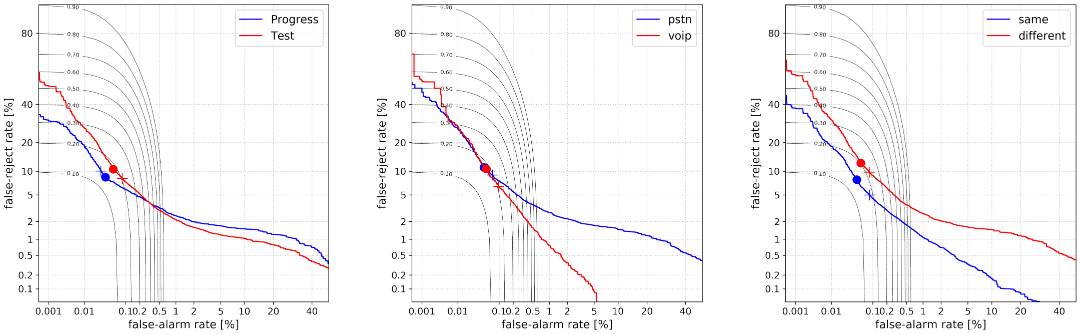


Figure 5: *Performance comparison of SRE18 vs SRE19 CTS submissions for several top performing systems.*

It is common practice in the machine learning community to perform statistical significance tests to facilitate a more meaningful cross-system performance comparison. Accordingly, to encourage the speaker recognition community to consider significance testing while comparing systems or performing model selection, we computed bootstrapping-based 95% confidence intervals using the approach described in [16]. To achieve this, we sampled, with repetition, the unique speaker model space along with the associated test segments 1000 times, which resulted in 1000 actual detection costs, based on which we calculated the quantiles corresponding to the 95% confidence margin. Figure 6 shows the performance confidence intervals (around the actual detection costs) for each submission for both the *progress* (top) and *test* (bottom) subsets. It can be seen that, in general, the *progress* subset exhibits a wider confidence margin than the *test* subset, which is expected because it has a relatively smaller number of trials. Also, notice that a majority of the top systems may perform comparably under different samplings of the trial space. Another interesting observation that can be made from the figure is that systems with larger error bars may be less robust than systems with roughly comparable performance but smaller error bars. For instance, although T_{18} achieves the lowest detection cost, it exhibits a much wider confidence margin compared to the second top system. These observations further highlight the importance of statistical significance tests while reporting performance results or in the model selection stage during system development, in particular when the number of trials is relatively small.

Figures 7a, 7b, and 7c show speaker recognition performance for the top performing submission in terms of DET curves as a function of: evaluation subset (i.e., *progress* vs *test*), CTS type (i.e., PSTN vs VoIP), and enrollment-test phone number match for PSTN calls (same vs different), respectively. The solid black curves in Figures 7a, 7b, and 7c represent equi-cost contours, meaning that all points on a given contour correspond to the same detection cost value. Firstly, consistent with our observations from Figure 4, the detection errors (i.e., false-alarm and false-reject errors) across the operating points of interest (i.e., the low false-alarm region) for the *test* subset are



(a) (b) (c)

Figure 7: *DET performance curves for the leading system by (a) data source (**progress** vs **test**), (b) CTS type (PSTN vs VoIP), and (c) enrollment-test phone number match (**same** vs **different**). Filled circles and crosses represent minimum and actual costs, respectively.*

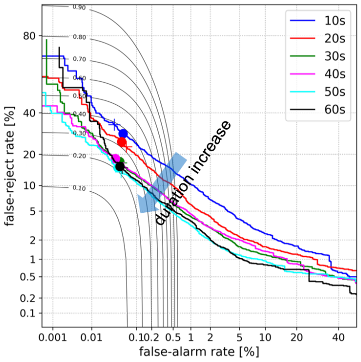


Figure 8: *DET curve performances of a top performing system for the various segment speech durations (10 s–60 s) in the test set.*

greater than those for the *progress* subset. In addition, the calibration error for the *test* subset is relatively larger. As noted previously, we speculate that these primarily result from overtuning/overfitting of the submission systems on the *progress* set. Secondly, contrary to the results observed on the SRE18 CTS domain where performance on the PSTN data was better than that on the VOIP data across all operating points, it seems from Figure 7b that for the operating points of interest (i.e., the low false-alarm region) the performance on the PSTN data is comparable to that on the VoIP data. We speculate this is due to the large amounts of VOIP data available for system development in SRE19 compared to SRE18 where only a small amount of VOIP development data was supplied. Finally, as expected, better performance is observed when speech segments from the same phone number are used in trials. Nevertheless, the error rates still remain relatively high even for the same phone number condition. This indicates that there are factors other than the channel (phone microphone) that may adversely impact speaker recognition performance. These include both intrinsic (variations in speaker's voice) and extrinsic (variations in background acoustic environment) variabilities.

Figure 8 shows DET curves for the various test segment speech durations (10 s–60 s) in the SRE19 CTS Challenge. Results are shown for a top performing submission. Limited performance difference is observed for durations longer than 40 s. However, there is a rapid drop in performance when the speech duration decreases from 30 s to 20 s, and similarly from 20 s to 10 s. This indicates that additional speech in the test recording helps improve the performance when the test segment speech duration is relatively short (below 30 seconds), but does not make a noticeable difference when there is at least 30 seconds of speech in the test segment. It is also worth noting that the calibration error (i.e., the gap between filled circles and crosses) increases as the test segment duration decreases.

7. Conclusion

In 2019, NIST organized the first leaderboard style SRE activity where raw CTS data (as opposed to embeddings) were provided as input to the systems. In this paper, we presented a summary of the SRE19 CTS Challenge (including the task, data, performance metric, the baseline system, as well as results and performance analyses) whose primary objectives were to systematically measure the recent progress in speaker recognition technology, in particular in the CTS domain, and to stimulate new ideas and collaborations. In addition, the CTS Challenge served as a prerequisite for the Audio-Visual SRE19. Results and analyses presented in this paper indicate great progress in speaker recognition technology compared to SRE18, with relative performance improvements as large as 70% for the leading system. Nevertheless, the performance gap on certain data partitions (e.g., PSTN vs VOIP or same vs different phone number) remains relatively large, at least for certain operating regions. This motivates further research towards developing a more robust technology that can maintain performance across a wide range of operating points and conditions (e.g., new data sources, languages, and channels).

8. Disclaimer

The results presented in this paper are not to be construed or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

9. References

[1] NIST, "NIST Speaker Recognition Evaluation," https://www.nist.gov/itl/iad/mig/speaker-recognition, [Online; accessed 28-December-2019].

[2] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, 2020.

[3] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proc. LREC*, Miyazaki, Japan, May 2018.

[4] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2019 NIST audio-visual speaker recognition evaluation," in *Proc. Speaker Odyssey (submitted)*, Tokyo, Japan, May 2020.

[5] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Graz, Austria, September 2019, pp. 1483–1487.

[6] D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, C. S. Greenberg, A. F. Martin, A. McCree, M. A. Przybocki, and D. A. Reynolds, "Summary and initial results of the 2013-2014 speaker recognition i-vector machine learning challenge," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 368–372.

[7] S. O. Sadjadi, "NIST baseline system for the 2019 speaker recognition evaluation CTS challenge," NIST, Tech. Rep., 2019.

[8] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 1353–1357.

[9] NIST, "Speech file manipulation software (SPHERE) package version 2.7," ftp://jaguar.ncsl.nist.gov/pub/sphere-2.7-20120312-1513.tar.bz2, 2012, [Online; accessed 28-December-2019].

[10] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, Rhodes, Greece, September 1997, pp. 1899–1903.

[11] NIST, "NIST 2019 Speaker Recognition Evaluation: CTS Challenge," https://www.nist.gov/document/2019nistspeakerrecognitionchallengev8pdf, 2019, [Online; accessed 27-December-2019].

[12] D. Povey *et al.*, "Kaldi Speech Recognition Toolkit," https://github.com/kaldi-asr/kaldi, [Online; accessed 01-March-2018].

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*. Calgary, AB: IEEE, April 2018, pp. 5329–5333.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. IEEE ICASSP*, May 2019, pp. 5796–5800.

[15] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proc. INTERSPEECH*, August 2011, pp. 2365–2368.

[16] N. Poh and S. Bengio, "Estimating the confidence interval of expected performance curve in biometric authentication using joint bootstrap," in *Proc. IEEE ICASSP*, vol. 2, April 2007, pp. II–137–II–140.

HALO – High Amplification Laser-pressure Optic

Alexandra Artusio-Glimpse, Kyle Rogers, Paul Williams, and John Lehman

National Institute of Standards and Technology, Boulder CO, USA
Corresponding e-mail address: alexandra.artusio-glimpse@nist.gov

Efforts are underway at the National Institute of Standards and Technology to drastically reduce the uncertainty of laser power measurements using radiation pressure. The High Amplification Laser-pressure Optic (HALO) system is a cornerstone of this effort as it enables amplification of the laser pressure on a high-quality mirror attached to a precision force sensor. We discuss the HALO architecture here.

MOTIVATION

Laser power measurements based on radiation pressure (RP) use the force of light as it reflects from a mirror to characterize the optical power. Here, we work to reduce measurement uncertainty by amplifying the force of the laser light. If the same light reflects multiple times from the sensing mirror of an RP detector, a passive gain can be realized [1,2]. For decades, the cryogenic radiometer standard has yielded relative measurement uncertainties of 10^{-4} at the milliwatt level. Meanwhile, power measurements above 100 W have relative uncertainties near 10^{-2} for both calorimetric-based and RP-based measurements. By amplifying the force of RP with a multi-reflection optical system, we expect laser power measurement uncertainties to approach 10^{-4} for kilowatt level incident powers.

For a RP measurement where the light reflects N times off the sensing mirror having reflectance R given a round-trip scattering and absorption loss L the signal to noise ratio (SNR) goes as

$$SNR \propto \sum_{j=1}^{N}\big((1-L)R\big)^{j-1}, \qquad (1)$$

which approaches N as $(1-L)R \rightarrow 1$. Absorption in the sensing mirror can heat the force sensor and produce a measurement error $N\eta_T\Delta T_1$ where η_T is the temperature-dependent error coefficient and ΔT_1 is the change in temperature of the force sensor for a single reflection. Thus, for low-loss, high-reflectivity optics, we approximate the fractional measurement error as

$$\varepsilon \approx (\eta_1/N + \eta_T\Delta T_1)/F_1, \qquad (2)$$

where η_1 represents the averaged noise and F_1 the measured force, both for a single reflection. We see that amplifying the force of a single reflection with N reflections, we effectively reduce the fractional contribution of the fixed noise by N, but have no effect on the thermal contribution.

Here we describe a system designed for $N \leq 15$ bounces and accommodating the 40 mm diameter beam from a 10 kW infrared laser. Current efforts emphasize the design and system tolerances, while forthcoming efforts will develop an alignment procedure and mirror reflectance and scatter measurements to meet preliminary tolerances. These are steps toward our goal of achieving relative measurement uncertainties approaching those currently found only in cryogenic radiometers. Moreover, our multi-reflection system can be used to amplify lower power lasers, enabling us to reduce the measurement uncertainty of any RP measurements.

HALO DESIGN

The NIST High Amplification Laser-pressure Optic (HALO) system is depicted in Fig. 1 with a green HeNe laser (5 mW) illuminating the beam path, reflecting off the sensing mirror 14 times, and leaving the system to the upper right.

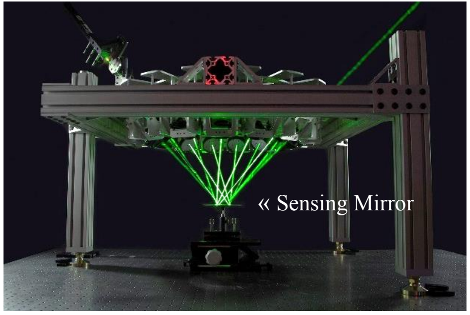


Figure 1. HALO system photograph. A green HeNe illuminates the laser path through 14-bounces and out the system to a beam dump (not shown) at the upper right.

HALO is a pentadecagonal structure with an entrance port and up to 14 upper mirror modules ("ring mirrors", see Fig. 2) that direct the input laser beam to a lower sensing mirror. Importantly, the laser incidence angle on the sensing mirror is always 45°. This both protects the mirror from thermal flexing as the reflectance need only be optimized for a single

Artusio-Glimpse, Aly; Rogers, Kyle; Williams, Paul A.; Lehman, John H. "HALO - High Amplification Laser-pressure Optic." Paper presented at NEWRAD 2020, 14th International Conference on New Developments and Applications in Optical Radiometry, Boulder, CO, US. June 23, 2020 - June 26, 2020.

angle and simplifies propagation of uncertainties. The total structure fills a volume of about 1 m^3. Like toroidal multipass cells used for laser spectroscopy [3], the laser beam traces out a star polygon (Fig. 2). However, here the ring mirrors are pitched downward to the sensing mirror placed at the center of the star pattern. When all 14 ring mirrors are in place, the total beam path in the system is 10.124 m.

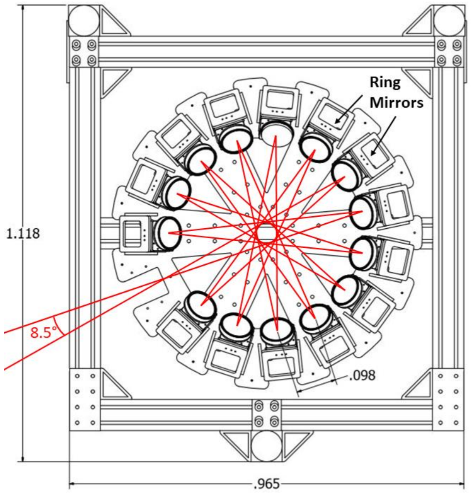


Figure 2. Upper ring mirror modules and star polygon laser path when all 14 mirrors (76.2 mm diameter) are in place. Laser enters through the empty port and exits through the same port, rotated by 8.5°. Alternatively, the laser may exit through any other port when the respective mirror module is removed. (labels in meters)

The spot pattern of the laser on the 150 mm diameter sensing mirror forms two open crescent shapes as the laser beam rotates and expands through the system (Fig. 3). Given the short (90 µm) coherence length of our laser and the large incidence angles and long path length differences between adjacent spots in this system, interference is not a concern.

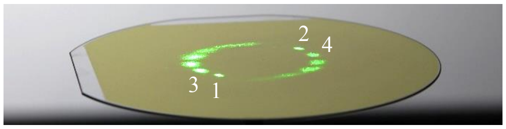


Figure 3. Green HeNe laser spot pattern is visible on the surface of a gold coated fused silica 150 mm diameter wafer. First four bounces of 14 total are labeled.

TIGHT TOLERANCES AND OUTLOOK

In order to reach 10^{-4} uncertainty, near perfect laser alignment and highly-accurate optical characterization of each mirror (reflectance, scatter, absorptance) will be required. As such, our first goal is to reach a relative measurement uncertainty of 10^{-3} at 10 kW.

We simulated laser propagation through the HALO system and modulated geometric and optical parameters to obtain system tolerances. The results of the ray-tracer and Monte Carlo algorithm are listed in Table 1 (totals exclude the force sensor [4]).

Table 1. Alignment and optical measurement tolerances needed to reach two relative uncertainty goals.

Rel. Unc. Goal	10^{-3}	10^{-4}
Sensing Mirror Angle	60 µrad	30 µrad
Ring Mirror Angle	300 µrad	30 µrad
All Mirrors 3D Position	2 mm	2 mm
Reflectance/Scatter	20x10^{-6}	1x10^{-6}
Total Optical Rel. Unc	0.7x10^{-3}	0.75x10^{-4}

In addition to tight system tolerances, if the sensing mirror is bowed, the laser will accumulate astigmatism, as is the case in Fig. 3 where the spots grow in ellipticity as the number of bounces increases. The HALO structure must also be rigid and isolated from vibrations to maintain these tolerances. To reach the 10^{-4} uncertainty goal, simply knowing total scattered power is not enough. The full bidirectional scatter distribution function must be accounted for in the momentum calculations else the calculated laser power from the measured force will be in error by an order of 10^{-4}.

The NIST HALO system has been built and its alignment demonstrated with a low-power, small diameter laser. We now set our attention to alignment with high reflectivity IR mirrors and measurement of a 500 W laser. This incremental advancement is necessary as we progress toward a RP laser power measurement system that will yield measurement uncertainties of 10^{-4} at 10 kW.

This publication is a contribution of the U.S. government and is not subject to copyright.

REFERENCES

1. S. Vasilyan, et al., Total momentum transfer produced by the photons of a multi-pass laser beam as an evident avenue for optical and mass metrology, Opt. Express, 25, 20798-20816, 2017.
2. G. Shaw, et al., Comparison of electrostatic and photon pressure force references at the nanonewton level, Metrologia, 56, 2019.
3. B. Tuzson, et al., Compact multipass optical cell for laser spectroscopy, Opt. Letters, 38, 257-259, 2013.
4. G. Shaw, et al., SI traceable electrostatic balance to measure laser power, NewRad 2020 extended abstract.

Artusio-Glimpse, Aly; Rogers, Kyle; Williams, Paul A.; Lehman, John H. "HALO - High Amplification Laser-pressure Optic." Paper presented at NEWRAD 2020, 14th International Conference on New Developments and Applications in Optical Radiometry, Boulder, CO, US. June 23, 2020 - June 26, 2020.

Ion-beam radiation damage to DNA by investigation of free radical formation and base damage

Melis Kant[a], Pawel Jaruga[a], Erdem Coskun[a], Samuel Ward[b], Alexander D. Stark[b], David Becker[b], Amitav Adhikary[b], Michael D. Sevilla[b] and Miral Dizdaroglu[a]
[a] Biomolecular Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899, USA
[b] Department of Chemistry, Oakland University, Rochester, MI 48309, USA

This work investigated the physicochemical processes and DNA base products involved in Ne-22 ion-beam (*ca*. 1.4 GeV) radiation damage to hydrated (12 waters/nucleotide) highly polymerized salmon sperm DNA. For this purpose, approximately 12 small (*ca.* 10 mm x 4 mm x 1 mm) samples were stacked in a sample packet and then ion-beam irradiated at 77 K. Free radicals trapped in ion-beam irradiated DNA at 77 K were elucidated using ESR spectroscopy. After warming the samples to room temperature, the measurement of DNA base damage by GC-MS/MS and LC-MS/MS with isotope dilution revealed the formation of a plethora of products from all four DNA bases, and also the formation of 8,5′-cyclopurine-2′-deoxynucleosides, namely 5′*R*-cyclo-dAdo, 5′*S*-cyclo-dAdo, 5′*R*-cyclo-dGuo and 5′*S*-cyclo-dGuo. This work is the first to use the combination of ESR and mass spectrometry, enabling a better understanding of the mechanisms of radiation damage to DNA along the ion-beam track in terms of the formation of DNA free radicals and products. ESR measurements showed that, as the linear energy transfer (LET) profile of ion-beam radiation increases, the production of cation, anion and neutral radicals of DNA increases along the ion-beam track. The yields of DNA damage products along the ion-beam track were in excellent agreement with the radical production. Because the probability of recombination of DNA radicals in the core increases due to the rise in concentration of proximate ion radicals, the location of the highest energy deposition, the Bragg peak, may show different damage and may not be the location of the maximum damage.

Supported by: NIH NCI grant R01CA04542, REF from CBR, OU, National Science Foundation under Grant No. CHE-1920110.

Kant, Melis; Jaruga, Pawel; Coskun, Erdem; Ward, Samuel; Stark, Alexander; Becker, David; Adhikary, Amitav; Sevilla, Michael; Dizdar, Miral M. "Ion-beam radiation damage to DNA by investigation of free radical formation and base damage." Paper presented at DNA Damage, Mutation and Cancer; Gordon Research Conference, Ventura, CA, US. March 01, 2020 - March 06, 2020.

Measurement of PARP1 in human tissues by liquid chromatography tandem mass spectrometry

Erdem Coskun [a,b], Gamze Tuna [a,c], Pawel Jaruga [a], Alessandro Tona [a], Onur Erdem [a,d], Miral Dizdaroglu [a]

[a] National Institute of Standards and Technology, Gaithersburg, MD, USA

[b] Institute for Bioscience & Biotechnology Research, University of Maryland, Rockville, MD, USA

[c] Department of Molecular Medicine, Institute of Health Sciences, Dokuz Eylul University, Izmir, Turkey

[d] Gulhane Faculty of Pharmacy, University of Health Sciences, Ankara, Turkey

Poly(ADP ribose) polymerase 1 (PARP1) is a multifunctional DNA repair protein of the base excision repair pathway and plays a major role in the repair of DNA strand breaks and in replication and transcriptional regulation among other functions. Mounting evidence points to the predictive and prognostic value of PARP1 expression in human cancers. Thus, PARP1 has become an important target in cancer therapy, leading to the development of inhibitors as anticancer drugs. In the past, PARP1 expression levels in tissue samples have generally been estimated by indirect and semi-quantitative immunohistochemical methods. Accurate measurement of PARP1 in normal tissues and malignant tumors of patients will be essential
for evaluating PARP1 as a predictive and prognostic biomarker in cancer and other diseases, and for the development and use of its inhibitors in cancer therapy. In this work, we present an approach involving liquid chromatography–isotope-dilution tandem mass spectrometry to positively identify and accurately quantify PARP1 in human tissues and cultured cells. We identified and quantified PARP1 in human normal ovarian tissues and malignant ovarian tumors, and in three pairs of human cell lines, each pair consisting of a normal cell line and its cancerous counterpart. Significantly greater expression of PARP1 was observed in malignant ovarian tissues than in normal ovarian tissues. In the case of one pair of cell lines, the cancerous cell line also exhibited greater expression of PARP1 than in normal cell line. We also show the simultaneous measurement of PARP1 and apurinic/apyrimidinic endonuclease 1 (APE1) in a given protein extract. The approach presented in this work is expected to contribute to the accurate quantitative assessment of PARP1 levels in basic research and clinical studies.

Coskun, Erdem; Tuna, Gamze; Jaruga, Pawel; Tona, Alessandro; Erdem, Onur; Dizdar, Miral M. ”Measurement of PARP1 in human tissues by liquid chromatography tandem mass spectrometry.” Paper presented at DNA Damage, Mutation and Cancer, Gordon Research Conference, Ventura, CA, US. March 01, 2020 - March 06, 2020.

Measurement of mass loss, absorbed energy, and time-resolved reflected power for laser powder bed fusion

David C. Deisenroth[a], Sergey Mekhontsev, Brandon Lane
National Institute of Standards and Technology[b], 100 Bureau Drive, Gaithersburg, MD 20899

ABSTRACT

Laser powder bed fusion processes are driven by scanned, focused laser beams. Along with selectively melting the metal powder, laser energy may be converted and transferred through physical mechanisms such as reflection from the metal surface, heat absorption into the substrate, vaporization, spatter, ejection of heated particles, and heating of the metal vapor/condensate plume that is generated by the laser-metal interaction. Reliable data on energy transfer can provide input for process modeling, as well as help to validate computational models. Additionally, some related process signatures can serve better process monitoring and optimization. Previous studies have shown that the proportion of the transfer mechanisms depend on laser power, spot size, and scan speed. In the current investigation, the energy conservation principle was used to validate our measurement of reflected energy, absorbed energy, and energy transfer by vaporization on bare plates of Nickel Alloy 625 (IN625). Reflected energy was measured using an optical integrating hemisphere, and heat absorbed into the substrate was measured by calorimetry. Transfer from vaporized mass loss was measured with a precision balance and used to establish an upper bound on energy transfer by mass transfer. In addition to measurement of total reflected energy, the reflected laser power was time-resolved at 50 kHz in the integrating hemisphere, which provided insight into the process dynamics of conduction, transition, and keyhole modes.

Keywords: Laser powder bed fusion, laser coupling, energy transfer in laser melting

1. INTRODUCTION

Metal laser powder bed fusion (LPBF) uses selective laser melting of metal powder to additively manufacture a part layer by layer. Complex physical phenomena occur at the laser-metal interaction area due to phase transitions, vapor pressure, surface tension gradients, wetting effects, melt dynamics, and more[1–3]. The energy density applied to the material has a strong effect on the physics that occur in the meltpool, and ultimately on the outcome of the laser melting process.

There are varying definitions of energy density reported in terms of energy divided by length, energy divided by volume, etc., but for the purposes of the current discussion, a strict definition of energy density is not necessary[4,5]. Regardless of definition, it is qualitatively known that the energy density is proportional to applied power and inversely proportional to scan velocity and beam spot size. Therefore, for a given spot size and scan velocity, decreased laser power decreases meltpool width and depth[6].

The energy density applied to the material is also related to the rate of metal vapor generation from the process. At process conditions producing little to no vapor jet, "conduction mode" occurs. Conduction mode is associated with meltpool aspect ratio (depth divided by ½ the width) of less than unity[7]. Among other defects, insufficient energy density is associated with lack of fusion and balling defects in LPBF[8]. At higher energy densities that result in a high rate of vapor generation, the process transitions into "keyhole mode," in which the meltpool depth increases substantially and the aspect ratio can become much greater than unity. A steep increase in laser power/energy absorption is associated with the transition from conduction mode to keyhole mode due to multiple reflections in the deep cavity formed in the meltpool by the vapor recoil pressure[9–11]. Among other defects, excessive or unstable keyholing is associated with residual porosity and loss of volatile alloy components[12,13].

The ability to measure the power and energy transfer that occur in this complex physical process enhances understanding of the process, aids in validation of computational models, and will ultimately help to improve the quality of parts built

[a] david.deisenroth@nist.gov; phone 1 301 975-2594

[b] This publication was prepared by United States Government employees as part of their official duties and is, therefore, a work of the U.S. Government and not subject to copyright.

Deisenroth, David; Mekhontsev, Sergey; Lane, Brandon. "Measurement of mass loss, absorbed energy, and time-resolved reflected power for laser powder bed fusion." Paper presented at SPIE Photonics West, San Francisco, CA, US. February 01, 2020 - February 06, 2020.

with LPBF. This preliminary investigation sought to measure the energy transfer of the laser melting process via absorption, reflection, and vaporization. Heat energy absorbed by the substrate was measured via calorimetry and evaporated energy was estimated by mass transfer. Reflected laser energy was measured as time-resolved reflected power at 50 kHz, which provided an additional wealth of information about the dynamics of the process in conduction, transition, and keyhole mode.

Measurement of light reflected from laser melting of metal is a very challenging application of reflectometry. The laser-metal interaction area is a highly dynamic reflective surface, which may potentially cause loss of reflected light from the reflectometer when reflected light propagates in the direction with a reduced throughput; the throughput is the ratio of the flux reaching the detector to the input flux from the source. The incident and reflected light can also be scattered and absorbed by the hot metal plume. At the current state of development, these two potentially important losses could not be quantified, although possible avenues for minimizing and/or measuring those effects are in progress. Nevertheless, the sum of the absorbed, reflected, and evaporated energy compared with the input laser energy places bounds on the combined effects of reflected light or energy absorption by the plume.

Laser power reflectometry can complement thermographic measurements and provide additional insight into the complex processes involved in building thin walls and overhangs. Full hemispherical capture of reflected light imposes significant limitations on the process laser incident angles and can find only a limited use in multilayer builds, which will be discussed in the following sections. As an additional benefit, reflected laser power also appears to be a highly useful process monitoring signature for defect detection[14].

2. EXPERIMENTAL METHODS

The experiments reported here were performed in the National Institute of Standards and Technology (NIST) Additive Manufacturing Metrology Testbed (AMMT)[15,16]. The AMMT is a custom LPBF research platform that was designed to be highly configurable for measurement of all aspects of the LPBF process. The AMMT includes a removable carriage that contains the build-well and a large metrology-well, both of which may be moved laterally within the large build chamber. The laser is an Yb-doped fiber laser with emission wavelength of 1070 nm. Laser power delivery can be adjusted from 20 W to more than 400 W, with a 4-sigma diameter (D4σ, representing diameter within which about 95 % of the Gaussian laser power profile is contained) spot size that is adjustable from 45 µm to more than 200 µm. The laser spot can be scanned with full control of the laser scan path/strategy at 100 kHz and laser power control at 50 kHz, with scan velocity from 0 mm/s to more than 4000 mm/s.

In the current investigation, all scans were performed with a velocity of 500 mm/s, a D4σ spot size of 65 µm, and laser power ranging from 50 W to 300W. The working material was rolled and annealed bare plates of Nickel Alloy 625 (IN625)[c] to avoid the measurement complications imposed by powder in this validation experiment. The manufacturer-reported composition of the Alloy 625 material used in this investigation are shown in Table 1.

Table 1. Constituents by mass % of IN625 used in this investigation

Element	Ni	Cr	Mo	Fe	Nb	Mn	Al	Ti
Mass %	60.6	21.98	8.4	4.38	3.44	0.35	0.21	0.21
Element	Si	Cu	C	Co	Ta	P	S	
Mass %	0.19	0.08	0	0.04	0.01	0.01	0.0001	

The surfaces of the plates were ground with 400 grit silicon carbide paper. The samples were assured to be within ±20 µm flatness and levelness of the build plane, resulting in an additional spot size uncertainty of ±1.3 µm due to the known caustic (solid angle of convergence). The following three experimental subsections will describe the approaches used for the three primary measurands: mass loss, laser energy absorption, and reflected laser power. It should be noted that the

[c] Certain commercial entities, equipment, or materials may be identified in this document to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

reflected power and absorbed energy experiments were done simultaneously, and the mass loss experiments were performed separately on a larger workpiece.

2.1 Mass loss

For measurement of mass loss caused by liquid metal spatter and/or vaporization, IN625 samples with ground surfaces were carefully cleaned with acetone and ethanol to remove debris. The samples were then stored in dry air for 24 hours to remove adsorbed water. The samples were transported in air-tight containers and then weighed on a precision balance. The laser melting process was performed with directional shield gas flow horizontal across the process at about 8 m/s, which was assumed to transport all of the spattered or evaporated material off of the 25 mm × 25 mm × 3 mm sample. After melting, the samples were carefully cleaned and dried again, then weighed.

The primary measurand was mass transfer, but an estimate of how much energy was transferred from the process due to mass loss can also be obtained. The upper bound of energy transfer due to the mass loss is found by assuming that all of the mass was evaporated. Any material that is evaporated absorbs heat energy through five modes in the transition from room temperature solid to superheated vapor, as shown in Equation (1).

$$E_{evap} \approx \Delta m[c_{p,sol}(T_m - T_r) + h_{fus} + c_{p,liq}(T_b - T_m) + h_{vap} + c_{p,vap}(T_{max} - T_b)] \quad (1)$$

Heat is first absorbed by the solid. The solid absorbs energy proportionally to the specific heat ($c_{p,sol}$) and the temperature rises from room temperature (T_r) to the to the melting temperature (T_m). Heat is then absorbed as the latent heat of fusion (h_{fus}) in the phase change from solid to liquid. The material is next heated as a liquid ($c_{p,liq}$) from the melting temperature to the boiling temperature (T_b). The largest energy transition is in the phase change from liquid to vapor (h_{vap}). Finally, the vapor is heated from the boiling temperature (T_b) to the maximum temperature the material reaches (T_{max}).

All values for the material heat capacity and latent heat were taken as the conservatively high values found in available literature. The solid specific heat is temperature dependent and was taken as the average value of about 0.6 J/g·K in the temperature range between room temperature and the melting temperature[17] of about 1300 °C. The latent heat of fusion is highest at the slowest cooling rates and was taken as the conservatively high value[18] of 290 J/g. Information on the heat capacities of IN625 is sparse above the melting point, but using the properties of the majority constituent (Ni) is a reasonable approximation[19]. The liquid latent heat[19] was taken to be 0.73 J/g·K up to Ni boiling point[20] of 2913 °C. Information on the specific heat of metal vapors is very sparse, so it was conservatively assumed that the vaporous specific heat was equivalent to that of the liquid. Precise temperatures of superheated vapor are also unknown, so it was conservatively assumed that the vapor temperatures reach up to 4500 °C. As stated previously, the phase transition from liquid to vapor is the largest energy conversion in the heating process; indeed, this single energy conversion amounts to about double the energy conversion of the solid-liquid phase change and other heating, combined.

Vaporization rates of liquid mixtures are dependent on the concentration of the constituents and the temperature difference between the vaporization temperature of each constituent and the hottest local material. Therefore, additional information is required to precisely calculate how much energy is converted by vaporization of IN625. Nevertheless, it is known that the latent heat of vaporization of the constituents, comprising about 85 % of IN625[21], is within ±5 % of 6.5 kJ/g. The remaining constituents are refractory and trace elements, and so it is a reasonable assumption that the mass loss can be converted to energy within a reasonable margin of uncertainty with this simplified approach. Uncertainty will be discussed in the next subsection.

2.2 Mass loss uncertainty

The balance used to measure the mass loss has a resolution of 0.01 mg and reproducibility of 0.015 mg for a combined mass uncertainty for a given measurement of 0.018 mg, and a mass change uncertainty of 0.025 mg. Estimating the uncertainty of the energy transfer due to the mass loss is far less straightforward, and given the assumptions, an conservative uncertainty of ±30 % was applied. It will be shown in the results section that the energy transferred to mass loss—with conservative assumptions and generous uncertainty assignment—was small compared to the uncertainty of the absorbed energy and reflected energy, precluding the need for further refinement of this energy transfer calculation. The next section will discuss the experimental methods used to measure energy absorbed into the workpiece.

2.3 Absorbed thermal energy

A small sample of about 3 mm × 3 mm × 8 mm was used for the simple calorimetry approach employed here. After surface preparation and attaching a calibrated thermistor, the sample was mounted on low thermal conductivity polymer pedestals. For low laser powers, longer and/or more scan lines were applied to the workpiece and in all cases a temperature rise of between 2.5 °C and 7.5 °C occurred in the workpiece. All scan tracks that were applied in immediate succession were spaced at least 300 µm apart to assure minimal thermal or metallurgical interaction between adjacent scans. The maximum scan time of 110 ms occurred at 50 W laser power. Three-dimensional finite element transient conduction simulations were performed to determine the time required for the workpiece to reach a temperature uniformity within that of the thermistor measurement uncertainty, which was a maximum of 10 s. The data after 10 s were then curve-fitted with Equation (2).

$$T(t) = T_0 + (T_e - T_0)e^{-t/\tau} \tag{2}$$

In Equation (2), $T(t)$ is the fitted temperature function of the thermistor temperature, T_0 is the initial temperature, T_e is the equivalent uniform initial temperature of the workpiece, t is time, and τ is the fitted curve time constant. All curve fits of the experimental data exhibited $R^2 > 0.9999$. Three-dimensional transient conduction simulations were also performed to assure that there was not an inordinate difference in average workpiece temperature decay between the highly non-uniform experimental surface temperature distribution in the initial 10 s and the ideal uniform workpiece surface temperature. The difference was found to be negligible compared to the measurement uncertainty. The equivalent uniform workpiece temperature (T_e) at 0 s was then used to find the absorbed laser energy (E_{abs}) with Equation (3), in which m is the workpiece mass and c_p is the workpiece specific heat.

$$E_{abs} = mc_p(T_e - T_0) \tag{3}$$

2.4 Absorbed thermal energy uncertainty

The uncertainty caused by the heat generation of the thermistor and the uncertainty of the mass of the sample were found to be negligible components of the uncertainty of the absorbed energy. The error in scan time is also believed to be negligible in the current investigation, leaving the remaining components to be those due to the uncertainty in specific heat and temperature rise of the workpiece. The temperature-dependent specific heat of IN625 was reported with standard uncertainty (1σ) of ±1.5 % by Pawel and Williams[22]. The uncertainty of the thermistor was 0.1 °C, resulting in a temperature difference uncertainty of 0.14 °C. These values were then incorporated into the standard propagation of uncertainty for each experiment. The next section will discuss the most challenging measurement of the current investigation, which was measurement of time-resolved reflected power.

2.5 Time-resolved reflected power

An ideal integrating sphere would capture all the light emanating from a source within the sphere, regardless of the direction, collimation, or transient nature of the light. Although an ideal integrating sphere is not possible in practice, appropriate provisions allow for highly efficient capture and measurement of the intensity of integrated light, regardless of how it is emanated. In the current application, the laser power reflected from the laser-melting process was measured.

The design of the integrating apparatus was constrained by the size of the build chamber, the necessary gas flow provisions for laser-metal interaction, and fabrication limitations. Use of a hemisphere instead of a full sphere allowed for a small laser-entrance port size and sample port size, relative to the total integration area, while fitting within the height of the build chamber of the AMMT. The integrating hemisphere was designed using as many best-practices as possible for isotropic throughput[23,24]. A cross-section view of a computer aided design model of the integrating hemisphere is shown in Figure 1, below. As shown, integration is facilitated by a diffuse, barium sulfate coating and with a specular, polished aluminum base. The focused laser light is directed through an elongated port on the top of the hemisphere, about 8° from vertical. The design allows scans to be done in an area of about 3 mm × 20 mm.

Another design constraint is the inert gas atmosphere required for laser melting to reduce detrimental oxidation[25]. Previous studies have shown that directional and inert shield gas flow is essential to facilitate continuous, consistent beam delivery by removing process biproducts that can distort, scatter, and obstruct beam delivery[26–28]. The base of the hemisphere was designed to be 6.7 mm thick due to an incorporated directional shield gas flow and oxygen sampling provision, shown in

Deisenroth, David; Mekhontsev, Sergey; Lane, Brandon. "Measurement of mass loss, absorbed energy, and time-resolved reflected power for laser powder bed fusion." Paper presented at SPIE Photonics West, San Francisco, CA, US. February 01, 2020 - February 06, 2020.

Figure 1. In the current investigation, laser scanning occurred in the direction perpendicular to the image plane of Figure 1, perpendicular to the directional shield gas flow.

Upon reaching the sample, some of the laser light is absorbed, causing melting, and some is reflected. The thickness of the base of the dome acts as a baffle and results in a reflected light capture angle of about 135°, as shown in Figure 1. The photodetectors (PD's) were located as close as possible to the equator of the hemisphere. At this location, the base thickness has a beneficial baffle-effect, and prevents deleterious direct viewing of reflected light by the PD's, which would circumvent integration. The base thickness, though, causes a loss of light in the remaining 45° of the hemisphere, which is not integrated. In practice, reflections from the laser-metal interaction at an angle less than 22° from the horizontal are unlikely, but possible, and the associated measurement uncertainty will be discussed in the uncertainty section. The fabricated base, hemisphere interior, and assembled integrating hemisphere are shown in Figure 2.

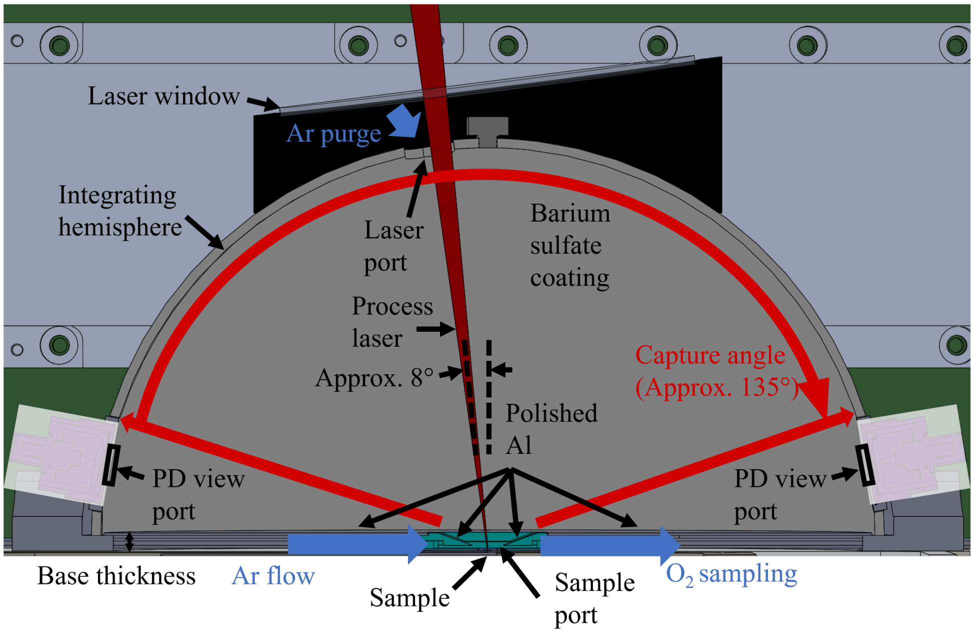


Figure 1. Cross-section view of a computer aided design model of the integrating hemisphere

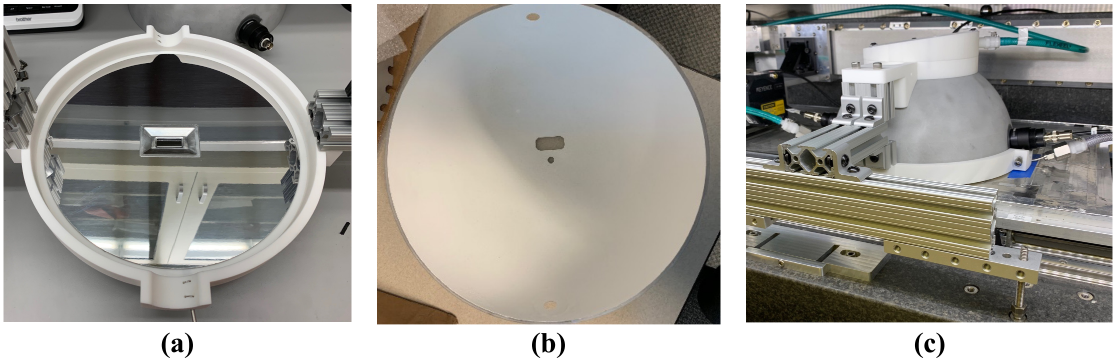


Figure 2. Images of the fabricated integrating hemisphere (a) specular base, (b) barium sulfate coated hemisphere interior, (c) assembled apparatus in build chamber.

The reflected power captured by the integrating hemisphere reflectometer was measured at a rate of 50 kHz. The PD signal was calibrated with a reflectance standard with reflectance of 0.98 at 1070 nm. The heating laser was defocused and directed onto the specular standard in order to correlate the PD signal with applied power. To convert the transient power measurements into total reflected energy from the process, the power was simply integrated in time, as shown in Equation (4).

$$E_{ref} = \int P_{ref}(t)\, dt \qquad (4)$$

2.6 Time-resolved reflected power uncertainty

The light loss due to the laser port and base thickness on the first reflection from the sample were measured by comparing the measured signal with a specular silver standard and a diffuse gold standard. On the first reflection, a specular sample does not result in any port loss or high angle loss. In contrast, a perfectly diffuse sample results in both high angle and port losses. After accounting for the reflectivity of the samples, it was found that the discrepancy was 2.5 % in intensity. Combined with the applied power (calibration) uncertainty of 2.5 %, the resulting reflected power uncertainty is 3.5 %. This value is taken as the estimated (Type B) standard uncertainty for the reflected power[29].

Throughput mapping of the inside of the hemisphere was performed to assure that, aside from the known losses, the relative location-dependent throughput of the integrating surface was uniform. The throughput is the ratio of the flux reaching the detector to the input flux from the source. Throughput mapping was performed with illumination around the equator of the hemisphere via 36 LED's at 850 nm wavelength. The reflectance of barium sulfate is expected to be similar in value at the LED and laser wavelengths. A photodetector was then placed at the entrance port and a gimbal-mounted mirror was mounted at the sample port. The mirror was aimed at a representative number of locations across the inside of the hemisphere. Although the mapping is not a perfect representation of the actual application, it is a good representation because of the similarity of reflectance at the LED and laser wavelengths and the known reciprocity of directional-hemispherical and hemispherical-directional measurements. Therefore, the measurements assure that there are no significant, unforeseen non-uniformities on the integrating surface.

As shown in Figure 3, the throughput uniformity of the surface is within ±1 % across the majority of the surface. It should be kept in mind that due to the Mercator projection; the corresponding true surface area is reduced by the cosine of the vertical angle. Thus, the entrance port area appears significantly greater than its true relative surface area. As was anticipated, significant losses and errors occur at a vertical angle of less than 22°.

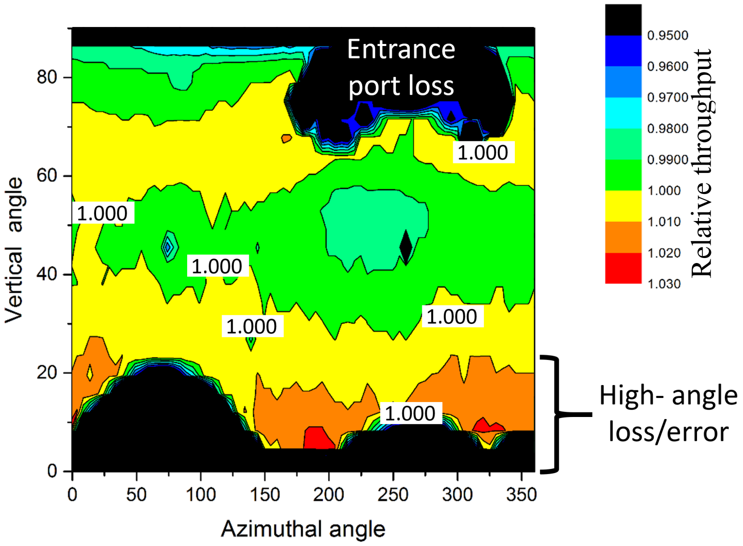


Figure 3. Relative throughput of the hemispherical integrating surface, shown in a Mercator projection.

At the current state of development, there are some significant unknowns about the character of the reflection from the laser-metal interaction. Depending on the applied energy density, the laser incident surface may be steeply inclined, and can reflect light forward along the scan direction, upward, or opposite to the scan direction[30,31]. The laser melting process is also highly dynamic, causing the reflective surface to rapidly change in time.

Further complicating matters, suspended vapor/condensate plume directly above the workpiece surface may also scatter and absorb both the incident and reflected laser light[32]. Incident laser light absorbed by the plume does not reach the workpiece and would not be measured by either calorimetry or reflectometry, which means that the fidelity of neither calorimetry nor reflectometry is compromised by incident laser light absorbed by the plume. Fidelity is retained because the energy reaching the workpiece is reduced and that would be accurately measured. However, plume absorption may affect the laser energy density applied to the surface, and thereby affect the melt pool formation and characteristics.

Incident laser light scattered by the plume may potentially be integrated and registered as light reflected from the workpiece, causing an erroneously high reflected light measurement. Laser light reflected from the workpiece that is then absorbed by the plume may cause an erroneously low measurement of the reflected power. Finally, laser light reflected from the workpiece that is then scattered by the plume would aid in integration, and therefore, not affect reflected power measurement. In summary, incident laser light scattering and reflected light absorption have counteracting erroneous effects on reflectometry, which are not yet well understood.

Therefore, with reflectometry alone, a significant portion of the reflected laser light could be lost or detected as erroneous reflection. Hence, the current investigation has sought to measure the sum of the measured reflected laser energy, the heat absorbed by the substrate, and the energy transfer to vaporization as a preliminary investigation into the magnitude of the combined measurement error and process energy loss to the plume.

The results of mass and energy transfer by vaporization or spatter, time-resolved reflected laser power, calorimetry, and combined energy balance will be presented in the following sections. It should be noted that the results shown in the "time-resolved reflected power" section were obtained with a more rudimentary experimental reflectometer setup than was described in this section, so the absolute values of normalized reflected power have greater uncertainty and port loss than the rest of the results. Nevertheless, the results are used to demonstrate the wealth of data that can be obtained with time-resolved non-contact reflectometry at 50 kHz.

3. RESULTS AND DISCUSSION

Examples of the melt tracks generated from the combined absorption and reflected power experiments are shown first, followed by the results of the mass and energy transfer due to vaporization. Then, time-resolved reflected power measurements with 20 mm long tracks with examples in conduction, transition, and keyhole mode are shown. And finally, the primary result of this effort, the combined energy balance obtained with the three forms of measurement, are discussed.

3.1 Melt tracks from reflectometer and calorimetry experiments

Examples of the melt tracks generated from the combined absorption and reflected power experiments are shown in Figure 4. The laser power at which the tracks were scanned are labeled. The tracks at each power were scanned from left to right and the track succession is from top to bottom. The lines at each laser power were scanned in immediate succession while the temperature of the workpiece was measured. In other words, each set of tracks labeled by applied laser power in Figure 4 represents an energy input that caused a temperature rise in the workpiece that was recorded until the workpiece cooled back to room temperature. It may be observed that the tracks of, for example, 50 W and 100 W are "interleaved," which was simply for convenience. Approximately equivalent energy (about 2 J) was input at each laser power, hence three tracks at 100 W were applied for equivalent energy input to a single 300 W track, and the three 50 W tracks are twice as long as the 100 W tracks. The track length ranged from 3 mm to 7 mm and the spacing between immediately successive tracks was at least 300 µm.

Cross-sections of the tracks formed for this study were not made, but the modes of laser-metal interaction are known from cross-sections performed by previous researchers under comparable conditions, as well as the characteristic changes in absorption[10]. The 60 W laser power resulted in conduction mode. At this power, the melt tracks were narrow and the solidified end of track shapes were short. In transition mode, at 100 W, the melt tracks are significantly wider and the solidified track ends were an elongated teardrop shape. At 300 W in keyhole mode, the melt track is the widest and the solidified track end was a highly elongated teardrop. In all cases, the melt tracks were visibly uniform and repeatable.

Deisenroth, David; Mekhontsev, Sergey; Lane, Brandon. "Measurement of mass loss, absorbed energy, and time-resolved reflected power for laser powder bed fusion." Paper presented at SPIE Photonics West, San Francisco, CA, US. February 01, 2020 - February 06, 2020.

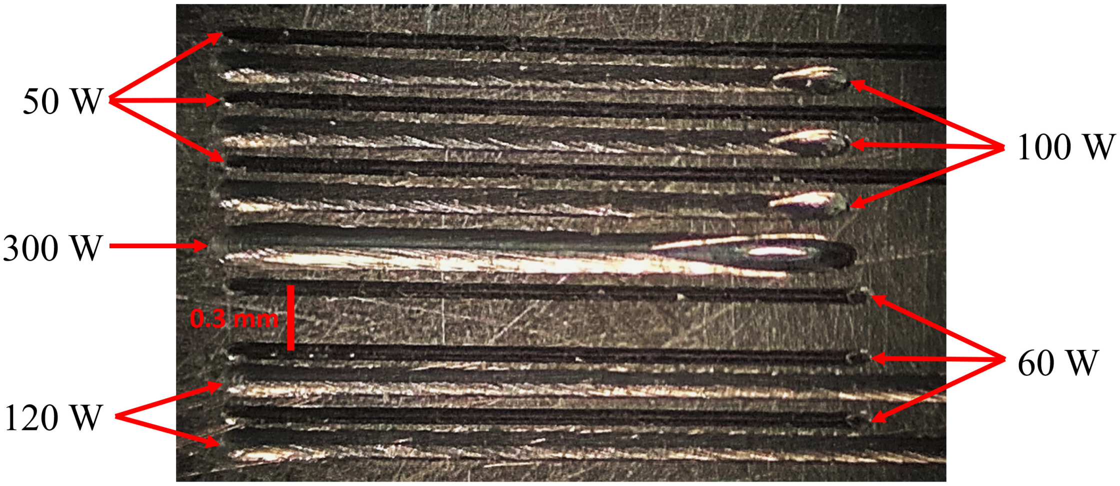


Figure 4. Melt tracks generated in the combined absorption and reflected power experiments with applied laser power pointing to tracks scanned in immediate succession

3.2 Mass and energy transfer due to vaporization

The measured mass transfer divided by distance for three laser powers and the estimated upper bound of normalized energy transfer from that mass are shown in Figure 5. Starting with 50 W, it can be observed that any mass loss was below the measurable limit, and therefore, energy transfer was also below the measurable limit. The mass loss divided by distance increased approximately linearly from no measured amount at 50 W, up to a value of 0.67 mg/m at 300 W.

The mass loss is about three times greater at 300 W compared to 100 W, causing the normalized estimated energy loss in transition mode and keyhole mode to be in close proximity. With the conservative estimate that all of the mass was transferred by vaporization and with an estimated uncertainty of 30 %, the total energy portion converted by mass loss was still less than 0.015 of the input energy. Therefore, because this transfer of energy was on the same order as the measurement uncertainty of both absorbed and reflected energy, energy loss due to mass transfer was taken as an additional 1.5 % uncertainty in the total energy balance for this preliminary investigation.

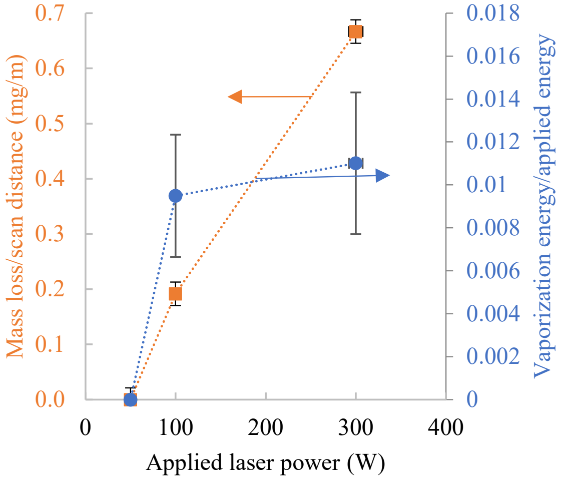


Figure 5. Mass transfer at varying laser power and estimated normalized energy transfer due to mass transfer

3.3 Time-resolved reflected power

The reflected power results were obtained with 20 mm long tracks under the same process parameters (spot size, scan speed, surface finish, etc.) as the rest of the results in this investigation, but with slightly greater uncertainty than the remainder of the results. Nevertheless, the results demonstrate the wealth of data that can be obtained with time-resolved non-contact reflectometry at 50 kHz, as shown in shown in Figure 6. In these results, three tracks were scanned in immediate succession and spaced 0.5 mm apart to avoid thermal and metallurgical interaction.

The reflected power from applied laser powers of 50 W, 100 W, and 200 W are shown in Figure 6 a, b, and c, respectively. Starting with Figure 6 a, about 0.50 of the laser power was reflected during each of the three scans. At the initiation of each 50 W track there is a slight overshoot with higher reflected laser power portion, up to 0.53 after the initial rise time. Overall, the conduction mode tracks exhibited very low variability along the length of the track.

In Figure 6 b, the 100 W tracks are in transition mode and show very high variability in reflected power along the length of the tracks. This high variability of reflected light in transition mode was observed in all experiments in this investigation and appears to be related to establishment of the vapor depression. The tracks exhibit nearly a 60 % difference in the reflected light portion at initiation of the track compared with the constant value of about 0.27 reached after the first 8 mm

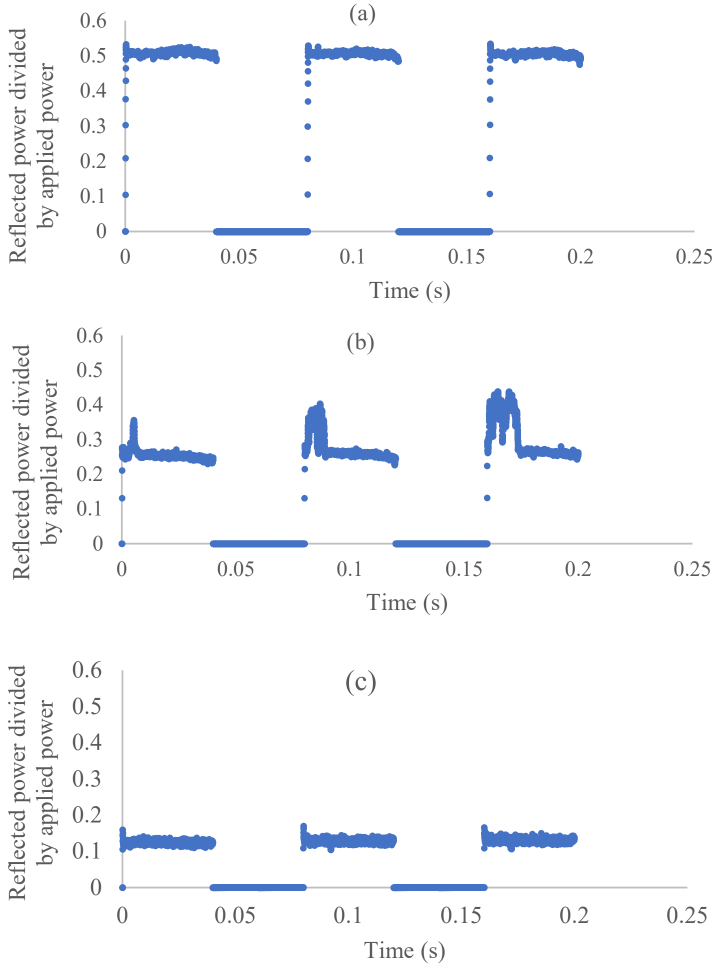


Figure 6. Time-resolved reflected power normalized by applied power. Each scan consists of three 20 mm long tracks scanned at (a) 50 W, (b) 100 W, and (c) 200 W

Deisenroth, David; Mekhontsev, Sergey; Lane, Brandon. "Measurement of mass loss, absorbed energy, and time-resolved reflected power for laser powder bed fusion." Paper presented at SPIE Photonics West, San Francisco, CA, US. February 01, 2020 - February 06, 2020.

of the track. It is currently unknown whether the apparent increase in track establishment length with each track is systematic or stochastic and will be investigated in the future.

Finally, in Figure 6 c, the 200 W tracks are in keyhole mode. At initiation of the track, the reflected light portion overshoots to 0.18 after the short initial rise time, then reaches an average value of about 0.125. In keyhole mode, the reflected light portion showed what appears to be a periodic fluctuation between 0.11 and 0.14 along the length of each scan. The frequency of such periodic fluctuations may be reported in the future.

3.4 Energy sum

The normalized reflected laser energy, absorbed heat energy, and their sum are shown as a function of applied laser power in Figure 7, with vaporization energy based on mass loss measurements added as a 1.5 % uncertainty component of the energy sum uncertainty. The energy values normalized by the applied laser energy indicate the relative proportion of reflected and absorbed energy, which would ideally sum to unity if all of the input energy were measured. It should be noted that the dotted lines in Figure 7 are not curve fits, but simply reference lines to guide the eye.

The trend and values of normalized absorbed energy are consistent with previous results under comparable conditions[9,10]. At 50 W laser power the absorbed energy portion is about 0.34, increasing to only 0.40 at 80 W. Then, the absorbed energy portion jumps to about 0.60 at 100 W and increases steadily to about 0.90 at 300 W. This jump in absorption at 100 W is associated with transition mode, while the low absorption is associated with conduction mode and the high absorption is associated with keyhole mode. The largest discrepancy between tests performed with the same laser power occurred in transition mode at 100 W, which appears to be related to the highly dynamic nature of absorption on the bare plates in transition mode, as was shown in Figure 6. Conversely to the portion of absorbed energy, the normalized reflected energy portion starts at 0.60 at 50 W, drops to about 0.30 at 100 W, and then the reflected energy portion falls to 0.07 at 300 W.

Turning now to the non-melting tests performed with an approximately 300 µm defocused laser spot under the same conditions, it can be seen from Figure 7 that a portion of 0.97 of the ideal energy sum was measured with 4.5 % uncertainty. The uncertainty is reported with 1σ confidence, indicating 68 % certainty that the measured value lies within the error

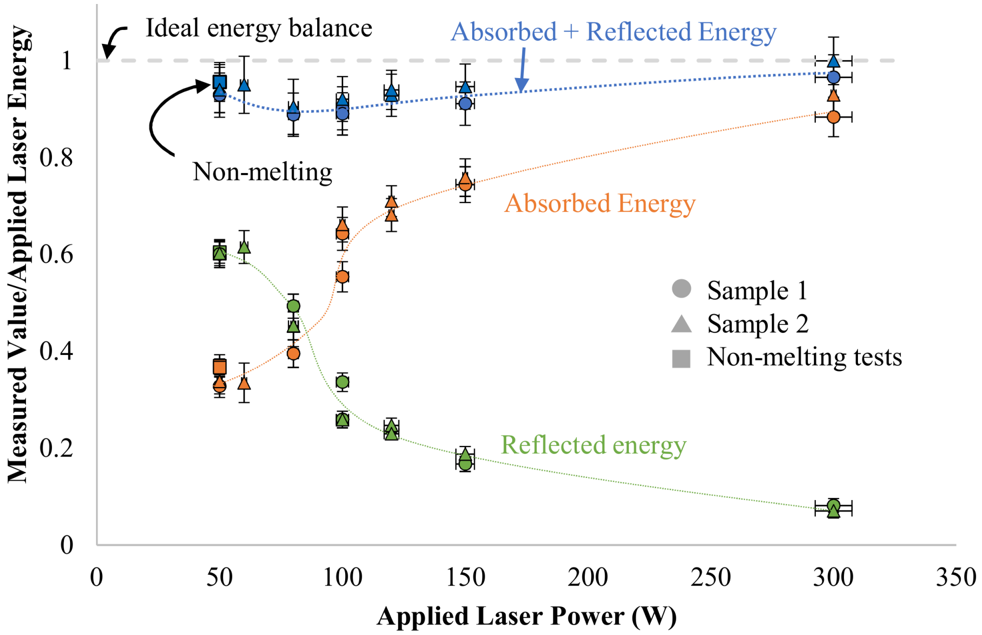


Figure 7. Reflected energy, absorbed energy, and the measured energy sum as a function of applied laser power. The dotted lines are to guide the eye along each data locus. Vertical and horizontal error bars indicate the combined standard uncertainty representing approximately 68 % confidence

bars. Therefore, the ideal energy balance was measured within the uncertainty under non-melting conditions, that likely generate a specular reflection.

In tests that caused melting, the error bars (which include energy transfer due to mass loss as an uncertainty) are within 1.5 % or less of the ideal energy balance, indicating that essentially all of the energy was measured at 50 W and 300 W. The maximum discrepancy between the ideal energy balance occurred at 80 W, at which between 7 % to 16 % of the applied energy was left unmeasured. In the current state of development, it is unclear whether this unmeasured energy is due to losses from the reflectometer (port losses or high-angle losses), or if a larger portion of the energy is converted to heating the ambient gas above the process in transition mode.

4. CONCLUSIONS

We are reporting the first known to us measurements of time-resolved reflected laser power from laser-matter interaction under conditions of interest for LBPF, which was accompanied by measurements of mass loss and absorbed laser energy. These three measurements performed together are meant to offer the modeling community a comprehensive set of data related to the laser-matter interaction processes. The measurement process can be streamlined so that such data sets can be developed for multiple materials and process parameters of interest.

In this investigation, tracks were laser melted into bare plates of IN625 with process conditions of interest for LPBF. The scan speed was 500 mm/s, the D4σ spot size was 65 µm, and the laser power was varied from 50 W to 300 W. Under the conditions tested, it was found that less than 1.5 % of the input energy was converted to vaporization. With time-resolved (50 kHz) reflected laser power measurements, it was observed that laser coupling is highly dynamic in transition mode, and periodic fluctuations in coupling were observed in keyhole mode.

Conservation of energy was confirmed to be measured in non-melting tests with a defocused beam. In tests with melting, conservation of energy was measured to be within 1.5 % or less of the 1σ error bars at the lowest and highest laser power. The maximum discrepancy between the measured and ideal energy balance occurred in transition mode, at which between 7 % to 16 % of the applied energy was left unmeasured. In the current state of development, it is unclear whether this unmeasured energy is due to port losses or high-angle losses in the integrating hemisphere, or if a larger portion of the energy is converted to heating the ambient gas/vapor/condensate above the process in transition mode.

Further developments depend on the interest of the community and can be directed either toward perfection of the reported measurements (such as reducing the uncertainties), expanding the functionality (e.g. measuring in the multilayer build condition with a powder layer), or connecting these measurands with the process outcomes.

ACKNOWLEDGEMENTS

The authors wish to thank to Mark Stoudt of the NIST Materials Measurement Laboratory for use of the precision balance for measuring mass loss. The authors also wish to thank you to Leonard Hanssen of NIST Physical Measurement Laboratory (PML) for aid in reflectometry uncertainty and analysis. The authors further wish to thank Brian Simonds of NIST PML for technical discussions. Finally, the authors with to thank Vladimir Khromchenko of NIST PML and Daniel Cardenas-Garcia of the National Metrology Center of Mexico for their work in characterizing the reflectometer.

REFERENCES

[1] Bidare, P., Bitharas, I., Ward, R. M., Attallah, M. M. and Moore, A. J., "Fluid and particle dynamics in laser powder bed fusion," Acta Materialia 142, 107–120 (2018).

[2] Wang, Y., Xing, L., Li, K., Yu, C., Ma, J., Liu, W. and Shen, Z., "Band-Like Distribution of Grains in Selective Laser Melting Track Under Keyhole Mode," Metallurgical and Materials Trans. B 50(2), 1035–1041 (2019).

[3] Keshavarzkermani, A., Marzbanrad, E., Esmaeilizadeh, R., Mahmoodkhani, Y., Ali, U., Enrique, P. D., Zhou, N. Y., Bonakdar, A. and Toyserkani, E., "An investigation into the effect of process parameters on melt pool geometry, cell spacing, and grain refinement during laser powder bed fusion," Optics & Laser Tech. 116, 83–91 (2019).

[4] Scipioni Bertoli, U., Wolfer, A. J., Matthews, M. J., Delplanque, J.-P. R. and Schoenung, J. M., “On the limitations of Volumetric Energy Density as a design parameter for Selective Laser Melting,” Materials & Design 113, 331–340 (2017).

[5] Fabbro, R., “Scaling laws for the laser welding process in keyhole mode,” J. Materials Processing Tech. 264, 346–351 (2019).

[6] Ghosh, S., Ma, L., Levine, L. E., Ricker, R. E., Stoudt, M. R., Heigel, J. C. and Guyer, J. E., “Single-track melt-pool measurements and microstructures in Inconel 625,” JOM 70(6), 1011–1016 (2018).

[7] Eagar, T. and Tsai, N., “Temperature-Fields Produced by Traveling Distributed Heat-Sources,” Weld. J. 62(12), S346–S355 (1983).

[8] Yadroitsau, I., [Selective laser melting: Direct manufacturing of 3D-objects by selective laser melting of metal powders], Lambert Academic Publishing (2009).

[9] Trapp, J., Rubenchik, A. M., Guss, G. and Matthews, M. J., “In situ absorptivity measurements of metallic powders during laser powder-bed fusion additive manufacturing,” Applied Materials Today 9, 341–349 (2017).

[10] Ye, J., Khairallah, S. A., Rubenchik, A. M., Crumb, M. F., Guss, G., Belak, J. and Matthews, M. J., “Energy Coupling Mechanisms and Scaling Behavior Associated with Laser Powder Bed Fusion Additive Manufacturing,” Adv. Eng. Materials, 1900185 (2019).

[11] Simonds, B. J., Sowards, J., Hadler, J., Pfeif, E., Wilthan, B., Tanner, J., Harris, C., Williams, P. and Lehman, J., “Time-Resolved Absorptance and Melt Pool Dynamics during Intense Laser Irradiation of a Metal,” Phys. Rev. Applied 10(4), 044061 (2018).

[12] Elmer, J. W., Vaja, J., Carlton, H. D. and Pong, R., “The effect of Ar and N2 shielding gas on laser weld porosity in steel, stainless steels, and nickel,” Weld. J. 94(10), 313s–325s (2015).

[13] Aboulkhair, N. T., Everitt, N. M., Ashcroft, I. and Tuck, C., “Reducing porosity in AlSi10Mg parts processed by selective laser melting,” Additive Manufacturing 1, 77–86 (2014).

[14] Coeck, S., Bisht, M., Plas, J. and Verbist, F., “Prediction of lack of fusion porosity in selective laser melting based on melt pool monitoring data,” Additive Manufacturing 25, 347–356 (2019).

[15] Lane, B., Mekhontsev, S., Grantham, S., Vlasea, M., Whiting, J., Yeung, H., Fox, J., Zarobila, C., Neira, J. and McGlauflin, M., “Design, developments, and results from the nist additive manufacturing metrology testbed (AMMT),” Proc. Solid Freeform Fabrication Symp., 1145–1160 (2016).

[16] Yeung, H., Neira, J., Lane, B., Fox, J. and Lopez, F., “Laser path planning and power control strategies for powder bed fusion systems,” Proc. Solid Freeform Fabrication Symp., 113–127 (2016).

[17] Haynes International., “HAYNES® 625 alloy: Principal Features,” H-3073F (2018).

[18] Tinoco, J. and Fredriksson, H., “Solidification of a Modified Inconel 625 Alloy under Different Cooling Rates,” High Temperature Materials and Processes 23(1), 13–24 (2011).

[19] Mills, K. C., [Recommended values of thermophysical properties for selected commercial alloys], Woodhead Publishing (2002).

[20] Lide, D. R., [CRC Handbook of Chemistry and Physics, 90th Edition], Taylor & Francis (2009).

[21] Zhang, Y., Evans, J. R. G. and Yang, S., “Corrected Values for Boiling Points and Enthalpies of Vaporization of Elements in Handbooks,” J. Chem. Eng. Data 56(2), 328–337 (2011).

[22] Pawel, R. E. and Williams, R. K., “Survey of physical property data for several alloys,” ORNL/TM--9616, Oak Ridge National Lab. (1985).

[23] Hanssen, L. M., Cagran, C. P., Prokhorov, A. V., Mekhontsev, S. N. and Khromchenko, V. B., “Use of a high-temperature integrating sphere reflectometer for surface-temperature measurements,” Int. J. Thermophysics 28(2), 566–580 (2007).

[24] Snail, K. A. and Hanssen, L. M., “Integrating sphere designs with isotropic throughput,” Applied Optics 28(10), 1793–1799 (1989).

[25] Ahn, J., He, E., Chen, L., Dear, J. and Davies, C., “The effect of Ar and He shielding gas on fibre laser weld shape and microstructure in AA 2024-T3,” J. Manufacturing Processes 29, 62–73 (2017).

[26] Malekipour, E. and El-Mounayri, H., “Common defects and contributing parameters in powder bed fusion AM process and their classification for online monitoring and control: a review,” Int. J. Adv. Manufacturing Tech. 95(1–4), 527–550 (2018).

[27] Ladewig, A., Schlick, G., Fisser, M., Schulze, V. and Glatzel, U., “Influence of the shielding gas flow on the removal of process by-products in the selective laser melting process,” Additive Manufacturing 10, 1–9 (2016).

[28] Anwar, A. B. and Pham, Q.-C., “Effect of inert gas flow velocity and unidirectional scanning on the formation and accumulation of spattered powder during selective laser melting,” Proc. of the 2nd Intl. Conf. on Progress in Additive Manufacturing, Singapore (2016).

[29] Taylor, B. N. and Kuyatt, C. E., “Guidelines for evaluating and expressing the uncertainty of NIST measurement results,” NIST Technical Note 1297, National Institute of Standards and Technology (1994).

[30] Zheng, H., Li, H., Lang, L., Gong, S. and Ge, Y., “Effects of scan speed on vapor plume behavior and spatter generation in laser powder bed fusion additive manufacturing,” J. Manufacturing Processes 36, 60–67 (2018).

[31] Ly, S., Rubenchik, A. M., Khairallah, S. A., Guss, G. and Matthews, M. J., “Metal vapor micro-jet controls material redistribution in laser powder bed fusion additive manufacturing,” Sci. Reports 7(1), 4085 (2017).

[32] Shcheglov, P. Y., Gumenyuk, A. V., Gornushkin, I. B., Rethmeier, M. and Petrovskiy, V. N., “Vapor–plasma plume investigation during high-power fiber laser welding,” Laser Phys. 23(1), 016001 (2012).

Anomalous accelerated negative-bias- instability (NBI) at low drain bias

K. P. Cheung
National Institute of Standards & Technology, Gaithersburg, MD USA

We observed at very low drain bias an anomalous acceleration of Negative-bias-instability at room temperature, as if the channel temperature has been raised significantly. The channel width and channel length dependent of this acceleration suggest that in addition to the conventional self-heating effect that raises the lattice temperature, there is indication that hot holes thermalized at a temperature (T_e) higher than the lattice temperature (T_L) is the cause of this anomaly. Analysis of the frequency dependent of drain bias modulation, as well as the wave shape dependent of the modulation support this explanation.

Self-heating accelerated channel-hot-carrier-degradation (CHD) and bias-temperature-instability (BTI) is a current concern for advanced technology nodes [1-3]. While there are disagreements on the severity of the problem, consensus is that there is an elevated lattice temperature (T_L) in the channel resulting from inhibition of thermal energy removal in advanced geometries such as FinFet or Gate-all-around (GAA) Fet, leading to accelerated CHD and BTI. The observed acceleration of degradations so far seems to entirely attributable to the increase in T_L, even though theoretical works [4, 5] suggest that non-equilibrium effect such as a carrier temperature T_e higher than T_L exists at the drain end of the channel and can lead to an even more severe degradation. This may be deal to most studies are done at stress conditions that produced strong "conventional" degradation that make it difficult to identify the non-equilibrium effect.

In this work, we choose a stress condition that should have weak expected conventional degradations to search for the non-equilibrium effects. From the thermal argument, a planar MOSFET (power devices excluded), particularly a standalone one in nanoscale, should not have notable self-heating accelerated degradation [1]. We choose normal NBTI gate bias but no elevated temperature to suppress normal NBTI. We choose low to very low drain bias to suppress CHD. Under these conditions, we observed a clearly anomalous acceleration of NBI. We argue that elevated electron temperature (T_e) may be the culprit. If true, this is perhaps the first example of reliability impact of high T_e. Naturally, we expect what we observed applies to advanced geometry as well.

The p-MOSFET used in our experiment are standalone 2μ x 50nm devices with 1.6nm (EOT) nitride gate oxide and polysilicon gate. All stresses are done using -2V gate bias. Stress-measure-stress sequence was done at fixed cycle: 600s stress, 4s wait, 5s measure, 4s wait, 5s measure, 4s wait, 5s measure. So, the total interval is 627 seconds. Checking the three measurements (V_t extraction: V_G sweep from -0.8V to 0V, V_D = -0.6V) found that they agree to within measurement noise (indicated by the error bars in fig. 1a). The average of the three is used as the measured result. All experiments were carried out at room temperature, apart from the standard NBTI reference run.

Fig. 1a shows the threshold voltage (V_t) shift as a function of stress time for various drain bias. The fitted lines have been forced to have zero intercept, meaning that V_t shift after 1 second stress is within the measurement noise. In other words, there is no sudden rise in V_t shift due to filling of existing traps as found in many reports in the literature. Fig. 1b shows the V_t shift as a function of drain bias after 7200 seconds of stress. A monotonic rise starting at zero bias is evident.

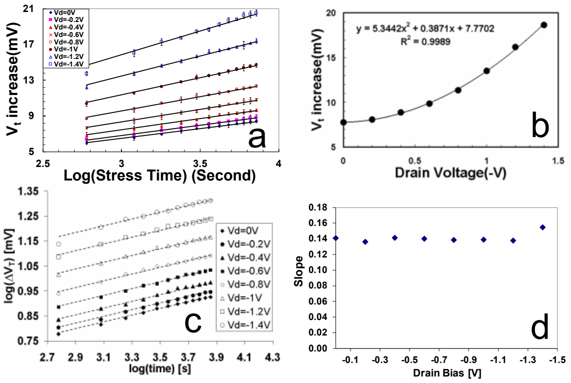


Fig.1 **a**. V_t shift as a function of stress time for various drain bias; **b**, V_t shift after 7200s stress as a function of drain bias; **c**, data in a replotted in log-log scale, **d**, slopes of the plots in c.

NBTI and CHD are intertwined and studies [2, 6, 7] showed that normally the overall degradation decreases with increasing drain bias until a turn-around point at which the increase in CHD overtakes the decrease in NBTI and degradation will increase with drain bias. The reason is at drain bias below -1V the population of energetic hot holes is so low that CHD is negligible. Since drain bias also reduces the oxide field in the channel, a reduction in NBTI as well as the overall degradation results. Contrary to these observations, our data showed only accelerated degradation even at very low drain bias. Note that our devices are from a production technology of a major foundry and very well behave. It is very unlikely that unusually bad devices are the culprit. Indeed, unusually bad devices that degrade easily even at very low drain bias means high density of preexisting defects. The fact that all lines cross zero (the fast, early degradation commonly reported is absent) in figure 1a indicates the quality of the devices studied here were superior to most of the devises reported in the literature. Thus, we call our result anomalous.

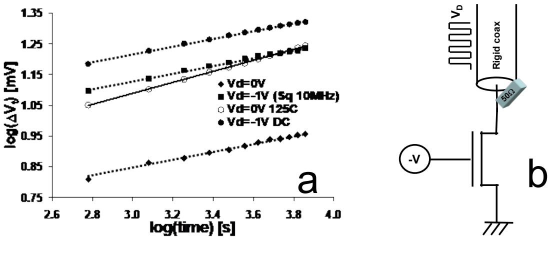


Fig. 2 **a**, V_t shift vs stress time for various conditions; **b**, measurement set up for high-speed drain modulation.

Adding to the anomaly is fig. 1c where the degradations are plotted in the more common log-log fashion and the slope of the fitted lines for all drain bias are 0.14 as shown in fig. 1d, lower than expected for NBTI or CHD [1-3]. To show that this is not an artefact of our experiment, a reference NBTI measurement (V_G = -2V, source/drain grounded, wafer heated to 125ºC using hot chuck) was performed and the slope was found to be 0.18 as

Cheung, Kin (Charles). "Anomalous accelerated negative-bias- instability (NBI) at low drain bias." Paper presented at 2020 International Reliability Physics Symposium, Virtual Dallas, TX, US. April 28, 2020 - May 30, 2020.

shown in fig. 2a. The 0.18 slope is consistent with the theoretical value when a finite delay is used in the measurement [1].

Also plotted in fig. 2a is a modulated rain bias case using a square wave at 10 MHz (0.5 ns rise/fall time). Note that this is not an AC stress with modulated gate bias. Instead, this is a "pulsed" drain bias. To ensure that the drain bias waveform is not distorted, a 50Ω terminated probe as shown in fig. 2b was used. This probe was verified to be good to 25 GHz using time-domain reflectometry measurement. As can be seen, the 10 MHz drain modulated case produced similar level of degradation as the 125°C zero bias case, as if the -1V modulated drain bias is heating the transistor to 125°C. However, the expected temperature rise is less than 10°C [1] for the 0.38 mW/μm power dissipation in this case. If CHD is indeed negligible as expected, then this large degradation is anomalous.

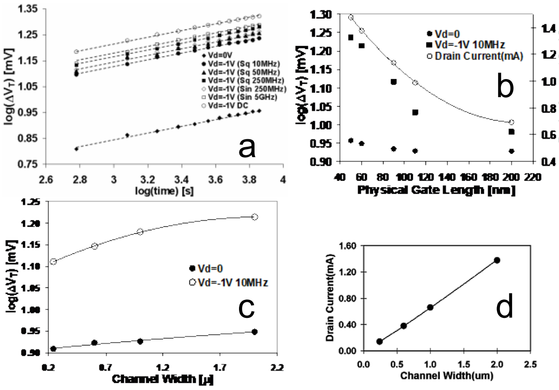


Fig. 3 **a**, V_t shift vs stress time for various drain bias frequency and wave shapes; **b**, V_t shift after 7200s stress and drain current vs channel lengths; **c**, V_t shift vs channel widths; **d**, drain current vs channel widths.

Fig. 3a shows more drain bias modulation cases involving different frequencies and wave shapes. For the square wave modulations, increasing the frequency from 10 MHz to 50 MHz to 250 MHz results in a commensurate increase in degradation, toward the level from DC bias. This again is counter to the expectation from heat removal argument [1] which says lower frequency modulation leads to higher peak temperature and therefore higher degradation. We note that if CHD were responsible for the degradation (instead of being negligible as we argued), the degradation could increase with frequency if gate voltage is modulated. Since gate voltage is steady and at least double the drain voltage at all time, this is not the case.

Sine waves were also used to modulate the drain bias at higher frequencies up to 5GHz (fig. 3a). We note that the degradation due to 250MHz sine wave modulation is very close to the degradation due to 50MHz square wave modulation. From this observation, we reason that the 5 GHz sine wave modulation should be like a 1GHz square wave modulation. The progression from 10MHZ to 1GHz square wave is asymptotically approaching the DC bias level.

Fig. 3b shows that log(degradation) tracks the drain current density (fixed channel width), suggesting that degradation increases exponentially with drain current density. Fig. 3C shows saturation of increases in log(degradation) with channel width. This means degradation increases with power dissipation even though the current density is unchanged (Drain current tracks the channel width as shown in fig. 3d).

Note that even though the degradations measured here are small, ranging from 8mV to 19mV, it is not insignificant given that the stress time is only 7200 seconds and the drain is only at operation voltage. For advanced nodes, 19mV is nearly 10% of V_t.

How to explain this collection of experimental results? We start with the channel width and channel length dependent behavior. The phonon mean-free-path for silicon is ~300nm [8] so the channel widths are all large enough that the thermal diffusion picture is valid (Fourier equation applicable). From narrow width to wide width the heat source changes from short line source to a long line source. For a uniform heat generation rate, the temperature will be higher for longer line. So, the channel width effect may be interpreted as conventional self-heating effect like those reported in the literature [1-3] (Similar to single fin vs multi fins in Finfets). Comparing fig. 3b and c, we see that the degradation change is significantly smaller for the channel width data set than the channel length data set. For the small power dissipation rate of 0.38mW/μm, we may assume that the thermal effect for the narrow width device is negligible and the entire range of degradation in the channel width data set is a measure of the conventional self-heating contribution, which translates to about 6mV out of the 19mV V_t shift.

Since the frequency dependent degradation counters the conventional self-heating picture, it further reinforces the notion that the larger degradation fraction is not due to conventional self-heating (T_L increase). In conventional self-heating heat diffusion is by acoustic phonons and T_e is the same as T_L. In nanoscale hot spot creation, local T_e and "temperature" of optical phonons (T_r) can rise much faster [9] and to a much higher level [4, 5], all riding on top of the slower responding conventional lattice temperature (T_L). How much higher than T_L obviously depends on both the drain current and drain bias. Both T_e and T_r provide a source of hot holes that can accelerate degradation. Also obvious is that at the very low drain bias level, both effects should be very weak. However, even a weak increase in hot holes population would impact the degradation rate.

With higher T_e and T_r the frequency dependent degradation can be explained as follows: As the drain bias is turned on, holes flow from source to drain. As they travel across the channel, they gain energy from the electric field and, at the same time, loss part of what they gain to phonon scattering. Upon arrival at the drain, some of these energetic carries lose their energy by creating optical phonons and the rest will thermalize extremely rapidly [9] to T_e by carrier-carrier scattering. The created optical phonons convert to acoustic phonons rapidly. Nevertheless, a steady state higher optical phonon population ($T_r > T_L$) can be expected.

A well-known argument [1] is that higher phonon scattering will lower the average carrier energy and therefore reduce the phonon generation rate. This argument will be important for the discussion later. For now, we point out that a T_e much higher than T_L will still result at steady state even with this negative feedback mechanism.

Since the optical phonon to acoustic phonon conversion is rapid, T_r follows the drain bias closely, meaning that $T_r \sim T_L$ at the end of the low drain bias cycle. When drain bias rises in the next cycle, T_r rises from a low value. Thus, every time there is a rising edge, there is a transient increase in T_e due to reduced optical phonon scattering as illustrated in fig. 4. Higher frequency means more rising edges, leading to more degradation. Eventually, frequency is high enough that T_r no longer falls with drain bias, one would approach the DC bias condition in agreement with data in fig. 3a.

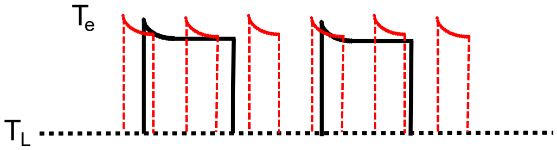


Fig 4. Profile of T_e and T_L in two frequencies (different color) showing the transient higher T_e at the rising edge.

What role does the T_L play in this scenario? T_L is too slow to follow the modulation at 10MHz or higher frequency. It is known that at low frequency, lower frequency leads to higher peak temperature [1]. That argument does not work here because at high frequency T_L does not respond fast enough. Any residual T_L modulation will be too small to show up as frequency dependent effect.

Substrate temperature is known to affect impact ionization rate in silicon and, at low drain bias the effect is positive, meaning that higher temperature leads to higher impact ionization rate [10]. (Note that, at high drain bias the opposite is true.) As impact ionization requires carrier energy higher than the drain voltage at low drain bias, the physics is very similar to what we are dealing with here – high energy carriers are needed to overcome the injection barrier into the SiO_2 layer for degradation to occur. It was suggested [10] that the high energy tail of the carrier distribution (at T_L) entering the channel preserves its shape after acceleration. This means higher T_L leads to higher T_e. Since even after acceleration the mean energy is still not high enough at low drain bias, it is the high energy tail of the T_e distribution that really matters. Consequently, higher T_L leads to higher degradation rate even when the dominant mechanism is T_e.

The reason why sine wave drain bias produces smaller degradation (~5x) than square wave can then be understood by the extremely fast thermalization of hot holes. It literally means that time spent at high voltage is what counts. Sine wave spend much less time at peak voltage.

Finally, what is the explanation for the low exponent of the degradation log-log plots? We do not have one at this point. Perhaps the fact that the channel is not uniformly heated may have something to do with it. However, we do notice that even when drain bias is zero at room temperature, the slope has the same low value.

In summary, we observed at very low drain bias an acceleration of Negative-bias-instability at room temperature, as if the channel temperature has been raised significantly. Many of the observed properties of the degradation is counter to the prevailing understanding of self-heating effects, thus are anomalous. The channel width and channel length dependent of this acceleration suggest that in addition to the conventional self-heating effect that raises the lattice temperature, there is indication that hot holes thermalized at a higher temperature is causing more degradation. Analysis of the dependence of drain bias modulation frequency, as well as the dependence on modulation wave shape further suggest non-equilibrium phonon distribution (optical phonon "temperature" higher than lattice temperature) also plays a role. The use of a regular planar MOSFET at low drain bias facilitated the study of the effect of carrier temperature because conventional self-heating effect is minimized, and CHD is also suppressed. The effect of high carrier temperature, if we are right, should exist in other geometries such as FinFets and GAA Fets.

Acknowledgement: Most of the data were collected by Dr. Yun Wang while he was a student at the Rutgers University.

[1] M. A. Alam et al, TED, 66(11), 4556-4565(2019).
[2] D. Son et al, EDL 40(9), 1354(2019).
[3] P. Paliwoda et al, TDMR 19(2), 249(2019).
[4] A. Abramo et al, JAP 76(10) 5786(1994).
[5] P. A. Childs et al, JAP 79(1), 222(1996).
[6] N. K. Jha et al, IRPS2005, pp524-528.
[7] P. Chaparala et al, Microelectr. Reliab. **45**(1), 13(2005).
[8] E. Pop et al, Proc. IEEE 94(8), 1587(2006).
[9] S. V. J. Narumanchi et al, Heat and Mass Transfer **V42**(6): 478-491(2006).
[10] B. Eitan et al, JAP 53(2), 1244(1982).

The 2019 NIST Audio-Visual Speaker Recognition Evaluation

Seyed Omid Sadjadi[1], Craig Greenberg[1], Elliot Singer[2,†], Douglas Reynolds[2,†], Lisa Mason[3], Jaime Hernandez-Cordero[3]

[1]NIST ITL/IAD/Multimodal Information Group, MD, USA
[2]MIT Lincoln Laboratory, MA, USA
[3]U.S. Department of Defense, MD, USA

craig.greenberg@nist.gov

Abstract

In 2019, the U.S. National Institute of Standards and Technology (NIST) conducted the most recent in an ongoing series of speaker recognition evaluations (SRE). There were two components to SRE19: 1) a leaderboard style Challenge using unexposed conversational telephone speech (CTS) data from the Call My Net 2 (CMN2) corpus, and 2) an Audio-Visual (AV) evaluation using video material extracted from the unexposed portions of the Video Annotation for Speech Technologies (VAST) corpus. This paper presents an overview of the Audio-Visual SRE19 activity including the task, the performance metric, data, and the evaluation protocol, results and system performance analyses. The Audio-Visual SRE19 was organized in a similar manner to the audio from video (AfV) track in SRE18, except it offered only the *open* training condition. In addition, instead of extracting and releasing only the AfV data, unexposed multimedia data from the VAST corpus was used to support the Audio-Visual SRE19. It featured two core evaluation tracks, namely audio only and audio-visual, as well as an optional visual only track. A total of 26 organizations (forming 14 teams) from academia and industry participated in the Audio-Visual SRE19 and submitted 102 valid system outputs. Evaluation results indicate: 1) notable performance improvements for the audio only speaker recognition task on the challenging *amateur* online video domain due to the use of more complex neural network architectures (e.g., ResNet) along with soft margin losses, 2) state-of-the-art speaker and face recognition technologies provide comparable person recognition performance on the *amateur* online video domain, and 3) audio-visual fusion results in remarkable performance gains (greater than 85% relative) over the audio only or visual only systems.

1. Introduction

The United States National Institute of Standards and Technology (NIST) organized the 2019 Speaker Recognition Evaluation (SRE19) in the summer–fall of 2019. It was the latest in the ongoing series of speaker recognition technology evaluations conducted by NIST since 1996 [1, 2]. The objectives of the evaluation series are 1) for NIST to effectively measure system-calibrated performance of the current state of technology, 2) to provide a common test bed that enables the research community to explore promising new ideas in speaker recognition, and 3) to support the community in their development of advanced technology incorporating these ideas.

SRE19 consisted of two separate activities: 1) a leaderboard-style Challenge using conversational telephone speech (CTS) extracted from the unexposed portions of the Call My Net 2 (CMN2) corpus collected by the Linguistic Data Consortium (LDC), which was also previously used to extract the SRE18 CTS development and test sets, and 2) a regular evaluation using audio-visual material extracted from the unexposed portions of the Video Annotation for Speech Technologies (VAST) corpus [3], also collected by the LDC. This paper presents an overview of the Audio-Visual SRE19 including the task, the performance metric, data, and the evaluation protocol as well as results and performance analyses of submissions. The SRE19 CTS Challenge overview and results are described in another paper [4]. It is worth noting here that the CTS challenge also served as a prerequisite for the Audio-Visual SRE19, meaning that in order to participate in the regular evaluation, one must have first completed the challenge (i.e., submitted to NIST valid system outputs along with sufficiently detailed system description reports). SRE19 was coordinated entirely online using a freshly designed web platform[1] deployed on Amazon Web Services (AWS)[2] that supported a variety of evaluation related services such as registration, data license agreement management, data distribution, system output submission and validation/scoring, and system description uploads.

The Audio-Visual SRE19 was organized in a similar manner to the audio from video (AfV) track of SRE18 [5], except it only offered the *open* training condition which allowed participants to use any publicly available and/or proprietary data for system training and development purposes. Moreover, in addition to the regular audio-only track, the Audio-Visual SRE19 also introduced audio-visual and visual-only tracks. Addition of these new tracks change the basic task in the Audio-Visual SRE19 to person detection (as opposed to speaker recognition), that is, determining whether a specified target person is present in a given test video recording. System submission was required for the audio and audio-visual tracks, but optional for the vi-

Table 1: *Audio-Visual SRE19 tracks.*

Track	Input	Required
Audio	Audio from Video	Yes
Audio-Visual	Audio and Frames from Video	Yes
Visual	Frames from Video	No

[†]The work of MIT Lincoln Laboratory is sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1] https://sre.nist.gov

[2] see Disclaimer.

Sadjadi, Omid; Greenberg, Craig; Singer, Elliot; Reynolds, Douglas; Mason, Lisa; Hernandez-Cordero, Jaime. "The 2019 NIST Audio-Visual Speaker Recognition Evaluation." Paper presented at The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, JP. November 01, 2020 - November 05, 2020.

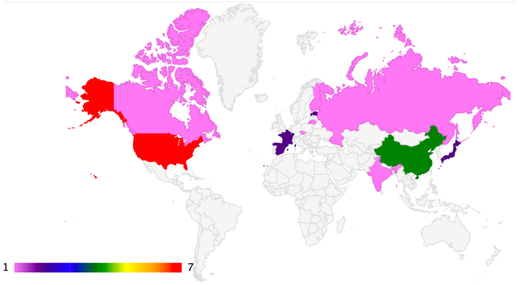


Figure 1: *Heat map of the world countries showing the number of Audio-Visual SRE19 participating sites per country.*

sual track. Table 1 summarizes the tracks for the Audio-Visual SRE19.

In addition, instead of extracting and releasing only the AfV data, unexposed multimedia data (i.e., videos) from the VAST corpus was used to support the Audio-Visual SRE19. Unlike the AfV track in SRE18 for which NIST released a very small in-domain development set containing data from only 10 speakers, SRE19 provided a much larger in-domain development set containing videos from 52 individuals from the VAST portion of SRE18 (i.e., only the videos in which the target individuals' faces were visible). In addition to the VAST development data, LDC also released selected data resources from the IARPA JANUS Benchmark-B [6], namely the JANUS Multimedia Dataset [7] which could also be used for system training and development purposes. The participants could register up to three systems for each track (i.e., audio, audio-visual, and visual), one of which under each track should have been designated as the primary system, and the other two as either contrastive or single best systems. Teams could make an unlimited number of submissions for each of the three systems until the evaluation period was over. Over the course of the evaluation, which ran from August 15, 2019 through October 21, 2019, a total of 14 teams, 8 of which were led by industrial institutions, from 26 sites made 102 valid submissions (note that the participants processed the data locally and submitted only the output of their systems to NIST for scoring and analysis purposes). Figure 1 displays a heatmap representing the number of participating sites per country. It should be noted that all participant information, including country, was self-reported. The number of submissions per team per track (i.e., audio, visual, and audio-visual) in the Audio-Visual SRE19 is shown in Figure 2.

Finally, as in SRE18, and in an effort to provide reproducible state-of-the-art baselines for the Audio-Visual SRE19, NIST released well in advance of the evaluation period a report [8] containing descriptions of speaker and face recognition baseline systems and results obtained using these standalone state-of-the-art (as of SRE18) deep neural network (DNN) embedding based systems as well as their fusion (see Section 5 for more details).

2. Task Description

The primary task for the Audio-Visual SRE19 was *person detection*, meaning that given a test video segment and a target individual's enrollment video, automatically determine whether the target person is present in the test segment. The test segment along with the enrollment segment from a designated target individual constitute a *trial*. The system is required to process each trial independently and to output a log-likelihood ratio (LLR), using natural (base e) logarithm, for that trial. The LLR for a given trial including a test segment s is defined as follows

$$LLR(s) = \log\left(\frac{P\left(s|H_0\right)}{P\left(s|H_1\right)}\right). \qquad (1)$$

where $P\left(\cdot\right)$ denotes the probability distribution function (pdf), and H_0 and H_1 represent the null (i.e., the target individual is present in s) and alternative (i.e., the target individual is not present in s) hypotheses, respectively.

3. Data

In this section we provide a brief description of the data released for the Audio-Visual SRE19 for system training, development, and test.

3.1. Training set

As noted previously, unlike in SRE18 which offered both *fixed* and *open* training conditions, the Audio-Visual SRE19 only offered the *open* training condition that allowed the use of any publicly available and/or proprietary data for system training and development purposes. The motivation behind this decision was twofold. First, results from the most recent NIST SREs (i.e., SRE16 [9] and SRE18) indicated limited performance improvements, if any, from unconstrained training compared to *fixed* training, although, participants had cited lack of time and/or resources during the evaluation period for not demonstrating significant improvement with *open* versus *fixed* training. Second, the number of publicly available large-scale data resources for speaker and face recognition has dramatically increased over the past few years (e.g., VoxCeleb[3]). Therefore, removing the *fixed* training condition would allow more in-depth exploration into the gains that could be achieved with the availability of unconstrained resources given the success of data-hungry Neural Network based approaches in the most recent evaluation (i.e. SRE18 [5]). Nevertheless, it is worth noting here that during the discussion sessions at the post-evaluation workshop, which was held in December 2019 in Singapore, several participating teams requested the re-introduction of the *fixed* training condition to facilitate meaningful and fair cross-system comparisons in terms of core speaker recognition algorithms/approaches (as opposed to particular data) used.

Although SRE19 allowed unconstrained system training and development, participating teams were required to provide a sufficient description of speech, non-speech (e.g., noise samples, room impulse responses, and filters), and visual data resources as well as pre-trained models used during the training and development of their systems.

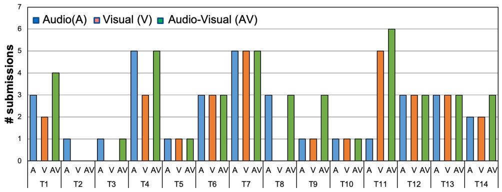


Figure 2: *Submission statistics for the Audio-Visual SRE19.*

[3] http://www.robots.ox.ac.uk/~vgg/data/voxceleb/

Table 2: *Statistics for the JANUS Multimedia Dataset (CORE) and the Audio-Visual SRE19 development (DEV) and TEST sets.*

Set	DEV/TEST	#speakers (M / F)	#Enroll segments	#Test segments	#Target	#Non-target
JANUS (CORE)	DEV	102*	102	319	244	32,294
	TEST	258*	258	914	681	235,131
SRE19 (AV)	DEV	15 / 37	52	108	108	5508
	TEST	47 / 102	149	452	452	66,896

*gender information not available

3.2. Development and test sets

For the sake of convenience, in particular for the audio-visual and visual-only tracks, NIST provided two *in-domain* development (DEV) sets that could be used for both system training and development purposes. The Audio-Visual SRE19 *DEV* sets were as follows:

- JANUS Multimedia Dataset (LDC2019E55)
- 2019 NIST Speaker Recognition Evaluation Audio-Visual Development Set (LDC2019E56)

The JANUS Multimedia Dataset (LDC2019E55) [7], which was extracted from the IARPA JANUS Benchmark-B datatset [6], was available from the LDC, subject to approval of the LDC data license agreement. It consists of two subsets, namely CORE and FULL, each with a *DEV* and *TEST* split. We only consider the CORE subset in this paper, because it better reflects the data conditions in the Audio-Visual SRE19 *DEV* and *TEST* sets where target speakers are assumed visible. The first two rows in Table 2 summarize the statistics for the JANUS Multimedia Dataset CORE subset.

The SRE19 Audio-Visual Development (*DEV*) Set (LDC2019E56), on the other hand, contained the original videos from which the VAST portion of the SRE18 *DEV* and *TEST* sets were compiled. Participants could obtain this dataset through the evaluation web platform (`https://sre.nist.gov`) after signing the LDC data license agreement. Unexposed portions of the VAST corpus were used to compile the Audio-Visual SRE19 *TEST* set. The second two rows in Table 2 summarize the statistics for the Audio-Visual SRE19 *DEV* and *TEST* sets.

The speech segments in the Audio-Visual SRE19 *DEV* and *TEST* sets were extracted from the VAST corpus collected by the LDC to support speech technology evaluations. Unlike existing publicly available datasets derived from online “red carpet” and interview style videos featuring celebrities (e.g., VoxCeleb[3]), the VAST corpus contains *amateur* video recordings such as video blogs (Vlogs) extracted from various online media hosting services. The videos are mostly shot using personal recording devices such as cell phones in extremely diverse acoustic backgrounds, illuminations, facial poses and expressions. The videos vary in duration from a few seconds to several minutes and include speech spoken in English. Each video may contain audio-visual data from potentially multiple individuals who may or may not be visible in the recording, therefore manually produced diarization labels (i.e., speaker time marks) and *keyframe* indices[4] along with bounding boxes that mark an individual’s face in the video were provided for both the *DEV* set and *TEST* set enrollment videos (but not for the test videos in either set). All video data were encoded as MPEG4. Figure 3 shows speech duration histograms for the enrollment and test segments in the Audio-Visual SRE19 *DEV* (left) and *TEST* (right) sets. Note that enrollment segment speech durations are calculated after applying diarization, while no diarization has been applied to test segments. Nevertheless, the enrollment and test histograms both appear to follow log-normal distributions, and overall they are consistent across the *DEV* and *TEST* sets.

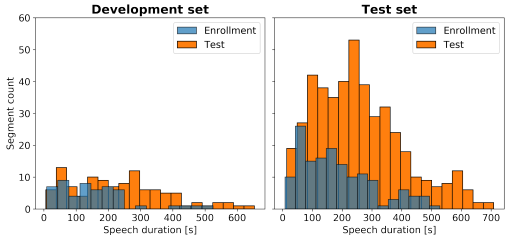


Figure 3: *Distributions of speech duration for the enrollment and test segments in the Audio-Visual SRE19 DEV and TEST sets.*

Similar to the AfV track in SRE18, there was only a 1-segment enrollment condition for the Audio-Visual SRE19 in which the system was given one video segment, that could vary in duration from a few seconds to several minutes, to build the model of the target individual. Note that for the audio track of the Audio-Visual SRE19, speech extracted from the enrollment video served as enrollment data, while for the visual track, face frame(s) (i.e., frames in which the face of the target individual was visible) extracted from the video served that purpose. Since NIST only released video files for the Audio-Visual SRE19, participants were responsible for extracting the relevant data (i.e., speech or face frames) for subsequent processing.

As in the most recent evaluations, gender labels were not provided for the enrollment segments in the *TEST* set. The test conditions for the SRE19 were as follows:

- The test segment video duration could vary from a few seconds to several minutes.
- The test video could contain audio-visual data from potentially multiple individuals.
- There were both same-gender and cross-gender trials.

4. Performance Measurement

Similar to past SREs, the primary performance measure for the Audio-Visual SRE19 was a detection cost defined as a weighted sum of false-reject (miss) and false-accept (false-alarm) error probabilities. Equation (2) specifies the Audio-Visual SRE19 primary normalized cost function for some decision threshold θ,

$$C_{norm}\left(\theta\right) = P_{miss}\left(\theta\right) + \beta \times P_{fa}\left(\theta\right), \qquad (2)$$

[4] Note that only a few (out of potentially many) target face frames per enrollment video were manually annotated.

where β is defined as

$$\beta = \frac{C_{fa}}{C_{miss}} \times \frac{1 - P_{target}}{P_{target}}. \quad (3)$$

The parameters C_{miss} and C_{fa} are the cost of a missed detection and cost of a false-alarm, respectively, and P_{target} is the *a priori* probability that the test segment speaker is the specified target speaker. The primary cost metric, $C_{primary}$ for the Audio-Visual evaluation was the normalized cost calculated at one operating point along the detection error trade-off (DET) curve [10], with $C_{miss} = C_{fa} = 1$, $P_{target} = 0.05$. Here, $\log(\beta)$ was applied as the detection threshold θ where log denotes the natural logarithm. Additional details can be found in the Audio-Visual SRE19 evaluation plan [11].

In addition to $C_{Primary}$, a minimum detection cost was also computed by using the detection threshold that minimized the detection cost.

5. Baseline systems

5.1. Speaker Recognition

In this section we describe the x-vector baseline speaker recognition system setup including speech and non-speech data used for training the system components as well as the hyper-parameter configurations used in our evaluations. Figure 4 shows a block diagram of the x-vector baseline system. The x-vector system is built using Kaldi [12] (for x-vector extractor training) and the NIST SLRE toolkit for back-end scoring.

5.1.1. Data

The x-vector baseline system was developed using the data recipe available at `https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2`. The x-vector extractor was trained entirely using speech data extracted from combined VoxCeleb 1 and 2 corpora. In order to increase the diversity of the acoustic conditions in the training set, a 5-fold augmentation strategy was used that added four corrupted copies of the original recordings to the training list. The recordings were corrupted by either digitally adding noise (i.e., babble, general noise, music) or convolving with simulated and measured room impulse responses (RIR). The noise and RIR samples are freely available from `http://www.openslr.org` (see [13] for more details).

5.1.2. Configuration

For speech parameterization, we extracted 30-dimensional MFCCs (including c0) from 25 ms frames every 10 ms using a 30-channel mel-scale filterbank spanning the frequency range 20 Hz–7600 Hz. Before dropping the non-speech frames using an energy based SAD, a short-time cepstral mean subtraction was applied over a 3-second sliding window.

For x-vector extraction, an extended TDNN with 12 hidden layers and rectified linear unit (RELU) non-linearities was trained to discriminate among the speakers in the training set. After training, embeddings were extracted from the 512-dimensional affine component of the 11th layer (i.e., the first segment-level layer). More details regarding the DNN architecture (e.g., the number of hidden units per layer) and the training process can be found in [14].

Prior to dimensionality reduction through LDA (to 250), 512-dimensional x-vectors were centered, whitened, and unit-length normalized. The centering and whitening statistics were computed using the in-domain development data (i.e., LDC2019E56). For backend scoring, a Gaussian PLDA model with a full-rank Eigenvoice subspace was trained using the x-vectors extracted from 170 k concatenated speech segments from the combined VoxCeleb sets as well as one corrupted version randomly selected from {babble, noise, music, reverb}. The PLDA parameters were then adapted to the in-domain development data (i.e., LDC2019E56) using Bayesian maximum *a posteriori* (MAP) estimation.

Finally, the PLDA verification scores were post-processed using an adaptive score normalization (AS-Norm) scheme proposed in [15]. We used LDC2019E56 as the cohort set, and selected the top 10% of sorted cohort scores for calculating the normalization statistics.

It is worth emphasizing that the configuration parameters employed to build the baseline system are commonly used by the speaker recognition community, and no attempt was made to tune the hyperparameters or data lists utilized to train the models.

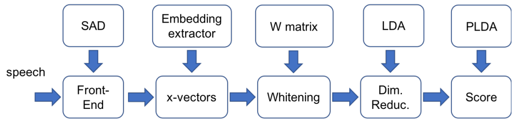


Figure 4: *A simplified block diagram of the baseline speaker recognition system for the Audio-Visual SRE19.*

5.2. Face Recognition

In this section, we describe the baseline face recognition system setup including the visual data used for training the system components as well as the hyper-parameter configurations used in our experiments. Figure 5 shows a block diagram of the baseline face recognition system which was built using open-source TensorFlow based implementations [16, 17] of 1) a face detector termed MultiTask Cascaded Convolutional Networks (MTCNN) [18], and 2) a face recognizer termed FaceNet [19] (for face encoding extraction). We use the NIST SLRE toolkit for back-end scoring.

5.2.1. Data

The baseline face recognition system utilized a pre-trained model available at `https://github.com/davidsandberg/facenet` (model name: 20180402-114759) which was trained on the VGGFace 2 dataset [20] using the Inception ResNet V1 architecture [21].

5.2.2. Configuration

We began processing by extracting one frame per second from the videos using `ffmpeg`. Then, we applied the MTCNN based face detector on the extracted frames to 1) filter out frames with no faces, and 2) compute the bounding box for the face

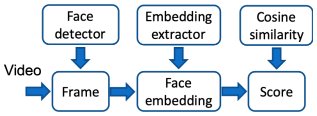


Figure 5: *A simplified block diagram of the baseline face recognition system for the Audio-Visual SRE19.*

that is closest to the center of the frame (as in [17]). Next, the face images were cropped using the bounding box coordinates, whitened (mean and variance normalized), and resized to 160 × 160 pixels. Finally, FaceNet was used to extract face encodings from the cropped, whitened and resized images.

For enrollment, we used the average of face encodings extracted from the enrollment video for each target individual to build a model for that individual. We only retained the face encodings that scored the highest (greater than 0.5 using cosine similarity) against the average of face encodings obtained using the manually produced bounding box coordinates for the enrollment videos. For test, we kept all face encodings extracted for each test video. In order to compute a single score for each trial involving an enrollment video and a test video, we computed the maximum of the cosine similarity scores obtained by comparing the enrollment encoding and test encodings. Finally, the scores were post-processed using the AS-Norm. We used the *DEV* set as the cohort set, and selected the top 10% of sorted cohort scores for calculating the normalization statistics.

6. Results and Discussion

In this section we present some key results and analyses for the Audio-Visual SRE19 submissions, in terms of the minimum and actual costs as well as DET performance curves.

Figure 6 shows the performance of the primary submissions per team per track, as well as performance of the baseline systems (see Section 5), in terms of the actual and minimum costs for the Audio-Visual SRE19 *TEST* set. Here, the y-axis limit is set to 0.5 to facilitate cross-system comparisons in the lower cost region. Several observations can be made from this figure. First, compared to the most recent SRE (i.e., SRE18), there seem to be notable improvements in audio only speaker recognition performance (see Figure 2b in [5]), which are largely attributed to the use of extended and more complex end-to-end neural network architectures (e.g., ResNet) along with soft margin loss functions (e.g., angular softmax) for speaker embedding extraction that can effectively exploit vast amounts of training data made available through data augmentation and/or large-scale datasets such as VoxCeleb[3]. Second, performance trends of the top 4 teams are generally similar, where the actual detection costs for the audio only submissions are larger than those for the visual only submissions, and the audio-visual fusion (i.e., the combination of speaker and face recognition system outputs) results in substantial gains in person recognition performance (i.e., greater than 85% relative in terms of the minimum detection cost for the leading system compared to their speaker- or face-recognition system alone). Third, more than half of the submissions outperform the baseline audio-visual system, with the leading system achieving larger than 90% improvement over the baseline. Fourth, in terms of calibration performance, mixed results are observed; for some teams (e.g., the top 2 teams) the calibration errors (i.e., the absolute different between the maximum and minimum costs) for speaker recognition systems are larger than those for face recognition systems, while for some others the opposite is true. Finally, in terms of the minimum detection cost, the two top performing speaker and face recognition systems achieve comparable results, which is a very promising outcome of this evaluation for the speaker recognition community, given the results reported in prior studies (e.g., see [7] where face recognition is shown to outperform speaker recognition by a large margin). It is worth emphasizing here that the top performing speaker and face recognition systems (i.e., team T_4) are both single systems (i.e., no fusion).

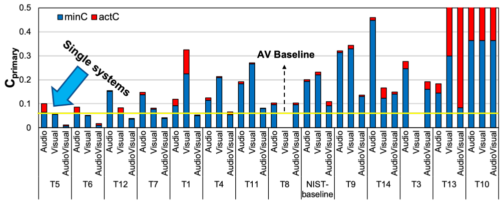


Figure 6: *Performance of the primary submissions for all three tracks (i.e., audio, visual, and audio-visual tracks) of the Audio-Visual SRE19 in terms of the minimum (in blue) and actual (in red) detection costs. The top performing audio and visual systems are both single systems (i.e., no fusion).*

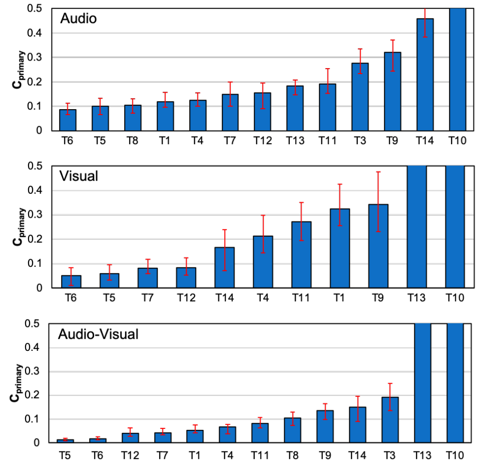


Figure 7: *Performance confidence intervals (95%) of the Audio-Visual SRE19 submissions for the audio (top), visual (middle), and audio-visual (bottom) tracks.*

It is common practice in the machine learning community to perform statistical significance tests to facilitate a more meaningful cross-system performance comparison. Accordingly, to encourage the speaker recognition community to consider significance testing while comparing systems or performing model selection, we computed bootstrapping-based 95% confidence intervals using the approach described in [22]. To achieve this, we sampled, with repetition, the unique speaker model space along with the associated test segments 1,000 times, which resulted in 1,000 actual detection costs, based on which we calculated the quantiles corresponding to the 95% confidence margin. Figure 7 shows the performance confidence intervals (around the actual detection costs) for each team for the audio (top), visual (middle), and audio-visual (bottom) tracks. It can be seen that, in general, the audio systems exhibit narrower confidence margins than their visual counterparts. This could be partly due to the fact that the majority of the participants, who are from the speaker recognition community, used off-the-shelf face recognition systems along with pretrained models not necessarily optimized for the task at hand in SRE19. Also, notice that several leading systems may perform comparably under different samplings of the trial space. An-

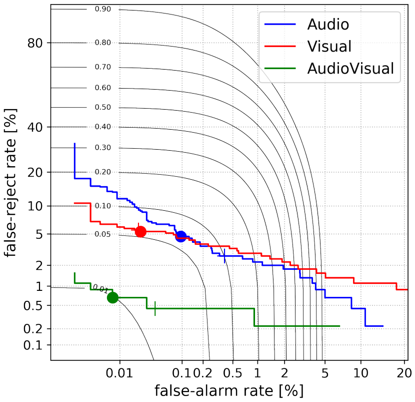


Figure 8: *DET curve performance of the top performing system for the* ***audio****,* ***visual****, and* ***audio-visual*** *tracks. Filled circles and crosses represent minimum and actual costs, respectively.*

other interesting observation that can be made from the figure is that audio-visual fusion seems to boost the decision making confidence of the systems by a significant margin, to the point where the two leading systems statistically significantly outperform the other systems. These observations further highlight the importance of statistical significance tests while reporting performance results or in the model selection stage during system development, in particular when the number of trials is relatively small.

Figure 8 shows DET performance curves from the leading system for the audio, visual, and audio-visual tracks. The solid black curves in the figure represent equi-cost contours, meaning that all points on a given contour correspond to the same detection cost value. Firstly, consistent with our observations from Figure 6 1) the audio-visual fusion provides remarkable improvements in performance across all operating points on the DET curve, which is expected given the complementarity of the two modalities (i.e., audio and visual), and 2) for a wide range of operating points, the speaker and face recognition systems provide comparable performance. Hence, the DET curves in Figure 8 confirm that the operating point dependent results in Figure 6 are consistent across a wider range of operating points, if not all of them.

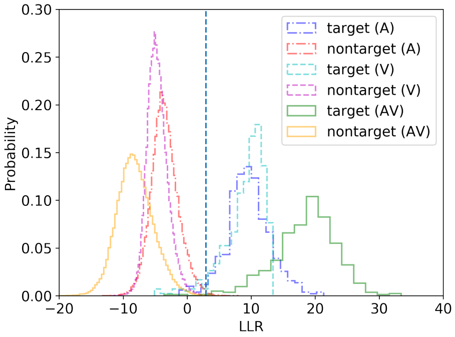


Figure 9: *Normalized target and non-target score distributions from the leading system for the* ***audio (A)****,* ***visual (V)****, and* ***audio-visual (AV)*** *tracks. The vertical dashed line represents the detection threshold.*

Motivated by the relatively low person recognition error rates achieved by the leading audio-visual system, i.e., 0.44% equal error rate (EER), we also conducted an error analysis of low scoring target and high scoring non-target trials, to gain insights regarding the nature of the issues associated with the remaining system errors on the Audio-Visual SRE19 *TEST* set. We found that, out of a total of 452 and 66,896 target and non-target trials, respectively, the system only made 2 false-reject (miss), and 27 false-accept (false-alarm) errors. A manual inspection of the trials (i.e., both enrollment and test videos) associated with these errors suggests that the majority of these trials indeed present challenging conditions for even humans (non-expert) due to the diversity of the acoustic backgrounds, illuminations, poses, facial expressions, and appearances (e.g., facial hair, glasses, caps/hats).

Figure 9 shows normalized target and non-target score distributions from the leading system for all tracks. The vertical dashed line represents the detection threshold. It can be seen that the score distributions from the audio only and face only systems roughly align, with the target and non-target distributions exhibiting some overlap at the threshold point. However, after the audio-visual fusion, the target and non-target classes are well separated with minimal overlap at the threshold, thereby significantly reducing the detection errors, in particular the false-rejects (misses).

7. Conclusion

Given the observed performance challenges presented by the AfV data in SRE18 and the growing interest of the speaker recognition research community in applying speaker recognition to more realistic multimedia applications, in 2019, NIST organized the first audio-visual SRE to 1) facilitate further exploration of speaker recognition technology in the AfV data domain, and 2) provide participants the opportunity to explore the possibility of fusing face and speaker recognition technologies. In this paper, we presented an overview of the Audio-Visual SRE19 activty including the task, data, the performance metric, the baseline system, as well as results and performance analyses. Compared to SRE18, the evaluation results indicate great progress in audio-only speaker recognition on the challenging AfV domain which is mainly attributed to the use of more complex neural network architectures (e.g., ResNet) along with soft margin losses. In addition, the audio-visual fusion was found to result in remarkable performance gains (greater than 85% relative) over the audio only or face only systems. Finally, state-of-the-art speaker and face recognition technologies were found to provide comparable person recognition performance on the challenging *amateur* online video domain.

8. Disclaimer

These results presented in this paper are not to be construed or represented as endorsements of any participant's system, methods, or commercial product, or as official findings on the part of NIST or the U.S. Government.

Certain commercial equipment, instruments, software, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

9. References

[1] NIST, "NIST Speaker Recognition Evaluation," https://www.nist.gov/itl/iad/mig/speaker-recognition, [Online; accessed 28-December-2019].

[2] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, 2020.

[3] J. Tracey and S. Strassel, "VAST: A corpus of video annotation for speech technologies," in *Proc. LREC*, Miyazaki, Japan, May 2018.

[4] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2019 NIST speaker recognition evaluation CTS challenge," in *Proc. Speaker Odyssey (submitted)*, Tokyo, Japan, May 2020.

[5] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, and J. Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Graz, Austria, September 2019, pp. 1483–1487.

[6] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother, "IARPA Janus benchmark-B face dataset," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 592–600.

[7] G. Sell, K. Duh, D. Snyder, D. Etter, and D. Garcia-Romero, "Audio-visual person recognition in multimedia data from the IARPA Janus program," in *Proc. IEEE ICASSP*, April 2018, pp. 3031–3035.

[8] S. O. Sadjadi, "NIST baseline systems for the 2019 audio-visual speaker recognition evaluation," NIST, Tech. Rep., 2019.

[9] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, Stockholm, Sweden, August 2017, pp. 1353–1357.

[10] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. A. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. EUROSPEECH*, Rhodes, Greece, September 1997, pp. 1899–1903.

[11] NIST, "NIST 2019 Speaker Recognition Evaluation Plan," https://www.nist.gov/document/2019nistmultimediaspeakerrecognitionevaluationplanv3pdf, 2019, [Online; accessed 27-December-2019].

[12] D. Povey *et al.*, "Kaldi Speech Recognition Toolkit," https://github.com/kaldi-asr/kaldi, [Online; accessed 01-March-2018].

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*. Calgary, AB: IEEE, April 2018, pp. 5329–5333.

[14] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *Proc. IEEE ICASSP*, May 2019, pp. 5796–5800.

[15] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proc. INTERSPEECH*, August 2011, pp. 2365–2368.

[16] I. de Paz Centeno, "MTCNN," https://github.com/ipazc/mtcnn, [Online; accessed 2-January-2020].

[17] D. Sandberg, "Face recognition using TensorFlow," https://github.com/davidsandberg/facenet, [Online; accessed 2-January-2020].

[18] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE CVPR*, June 2015, pp. 815–823.

[20] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, May 2018, pp. 67–74.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, pp. 4278–4284.

[22] N. Poh and S. Bengio, "Estimating the confidence interval of expected performance curve in biometric authentication using joint bootstrap," in *Proc. IEEE ICASSP*, vol. 2, April 2007, pp. II–137–II–140.

Effectiveness of dataset reduction in testing machine learning algorithms

Jaganmohan Chandrasekaran, Huadong Feng
Department of Computer Science & Engineering
The University of Texas at Arlington
Arlington, USA
{jaganmohan.chandrasekaran, huadong.feng}@mavs.uta.edu

Yu Lei
Department of Computer Science & Engineering
The University of Texas at Arlington
Arlington, USA
ylei@cse.uta.edu

Raghu Kacker, D. Richard Kuhn
Information Technology Lab
National Institute of Standards and Technology
Gaithersburg, USA
{raghu.kacker, d.khun}@nist.gov

***Abstract*— Many machine learning algorithms examine large amounts of data to discover insights from hidden patterns. Testing these algorithms can be expensive and time-consuming. There is a need to speed up the testing process, especially in an agile development process, where testing is frequently performed. One approach is to replace big datasets with smaller datasets produced by random sampling. In this paper, we report a set of experiments that are designed to evaluate the effectiveness of using reduced datasets produced by random sampling for testing machine learning algorithms. In our experiments, we use as subject programs four supervised learning algorithms from the Waikato Environment for Knowledge Analysis (WEKA). We identify five datasets from Kaggle.com to run with the four learning algorithms. For each dataset, we generate reduced datasets of different sizes using two random sampling strategies, i.e., pure random and stratified random sampling. We execute our subject programs with the original and the reduced datasets, and measure test effectiveness using branch and mutation coverage. Our results indicate that in most cases, reduced datasets of even very small sizes can achieve the same or similar coverage achieved by the original dataset. Furthermore, our results indicate that reduced datasets produced by the two sample strategies do not differ significantly, and branch coverage correlates with mutation coverage.**

***Keywords*— Testing classifiers, Random sampling, Reduced datasets, Testing machine learning, Branch coverage, Software testing.**

I. Introduction

Many machine learning algorithms examine large amounts of data to discover insights from hidden patterns. Testing machine learning algorithms can be expensive and time-consuming. There is a need to speed up the testing process, especially in an agile development process, where testing is frequently performed. One approach is to replace big datasets with smaller datasets produced by random sampling. One natural question to ask about this approach is the following: How does a reduced dataset compare to the original dataset in terms of effectiveness from a testing perspective?

In this paper, we report a set of experiments that are designed to evaluate the effectiveness of using reduced datasets for testing machine learning algorithms. We measure test effectiveness using both branch coverage and mutation coverage. In our experiments, we use as our subject programs four supervised learning algorithms from the Waikato Environment for Knowledge Analysis (WEKA), which is a widely used machine learning workbench written in Java [1]. We identify five datasets, each of which represents a different application domain, from Kaggle.com to run with these algorithms. Kaggle.com is an online data science community that maintains a repository of public datasets.

After we identify subject programs and datasets, we first execute each subject program with each of the five datasets and measure test effectiveness in terms of branch coverage and mutation coverage. Second, we create two groups of reduced datasets. The first group is generated using pure random sampling, i.e., in a purely random manner. The second group is generated using stratified random sampling, i.e., in a random manner that maintains the class distribution. In another word, a reduced dataset produced by stratified random sampling has the same class distribution as the original dataset. In the rest of the paper, we will refer to pure random sampling as random sampling and stratified random sampling as stratified sampling. Third, we execute the reduced datasets with subject programs and measure branch and mutation coverage. Finally, we compare the coverage results achieved by the reduced datasets to the coverage results achieved by the original datasets.

The major findings from our experiments are summarized as follows:

- In most cases, reduced datasets of even very small sizes achieve coverage identical or similar to the original datasets. In our experiments, the original datasets have the number of instances ranging from 142,193 to 999,999. The reduced datasets are of four sizes, i.e., 100, 200, 400, and 800, which are a fraction of the original dataset size. However, 522 out of 800 reduced datasets achieved the same coverage as the original datasets. Also, 112 out of 800 reduced datasets achieved more than 90% of the coverage achieved by the original datasets.
- One might expect that stratified sampling can be more effective than random sampling. However, in our experiments, the coverage results of the reduced datasets produced by the two sampling strategies are very similar. In particular, 628 out of 800 reduced datasets produced by the two sampling strategies achieved the same coverage. It is interesting to note that in several cases, random sampling achieved higher coverage than stratified sampling. The reason is that when the sample size is small, and when the dataset is skewed in terms of class distribution, stratified sampling may produce no instances for a particular class, which could significantly reduce coverage.

Kacker, Raghu N.; Kuhn, D. Richard. "Effectiveness of dataset reduction in testing machine learning algorithms." Paper presented at IEEE International Conference on Artificial Intelligence Testing, Oxford, UK. April 13, 2020 - April 16, 2020.

- In most cases, branch coverage correlates with mutation coverage. Since mutation testing is quite expensive to perform, this suggests that branch coverage could be used as a practical alternative in place of mutation coverage for testing machine learning algorithms.

The rest of the paper is organized as follows. In Section II we present the design of our experiments, including the research questions, subject programs, datasets and metrics used in our experiments, and discussion about the generation of reduced datasets. Section III presents the results of our experiments, including branch and mutation coverage results for original and reduced datasets as well as implications of these results. Section IV discusses potential threats to validity, including both internal and external threats. Section V reviews existing work that is related to ours. Section VI provides conclusion remarks and a few directions for the future work

II. Experimental Design

In this section, we present how we design our experiment, including the research questions, the selection of subject programs and datasets, the sampling approaches used to generate reduced datasets, and the metrics used to measure the effectiveness of the dataset executions.

A. Research Questions

Our main objective is to investigate the effectiveness of using a reduced dataset (in terms of volume, i.e., number of instances in a dataset) to test machine learning algorithms. We formulate the following research questions:

- How effective is it to test machine learning algorithms using reduced datasets, in comparison with the original datasets?
- How do the two sampling strategies, i.e., random sampling and stratified sampling, compare to each other?
- In testing machine learning algorithms, can branch coverage be used as a substitute for mutation coverage?

B. Subject Programs

Waikato Environment for Knowledge Analysis (WEKA) is a machine learning workbench developed by University of Waikato. WEKA has a collection of supervised and unsupervised algorithms implemented in Java. Using WEKA, a user can perform tasks such as classification, regression, clustering and association rule mining. Four supervised algorithms from WEKA are used as our subject programs.

In WEKA, classification algorithms are categorized into seven different groups. We select one algorithm from each of the following four groups, *bayes*, *meta*, *rules* and *trees.* When we choose one algorithm from a group, we only consider algorithms that satisfy two conditions: (1) they support datasets with nominal class labels and (2) they generate a *model* at the end of its training phase. When there are multiple algorithms that satisfy the two conditions, we randomly choose one from these algorithms. The reason for condition (1) is that we use WEKA`s built-in filter to generate smaller datasets for stratified sampling. This filter is applicable only to datasets with nominal class labels. The reason for condition (2) is that during mutation testing, we need expected output to determine if a mutant is killed by comparing against the actual output. If an algorithm generates a model, then the model can be used as expected output during mutation testing.

For example, WEKA lists eight algorithms under the *trees* category. However, one of the eight algorithms, M5p, does not work on a nominal class label. Hence we exclude M5P. Similarly, of the remaining seven algorithms, *random forest* works on a nominal class label dataset, but at the end of its training phase, the model is not accessible to the user with default configuration options. Hence, we exclude *random forest*. From the remaining six algorithms, we randomly select j48 as one of our subject algorithms.

Among different categories of classifiers listed in WEKA, we selected four algorithms namely NaiveBayes classifier [27], AdaBoost1 classifier [28], OneR classifier [29] and J48 classifier [30]. Table I lists our subject algorithms and some information about these algorithms, including package/class information, and number of branches. Each algorithm is executed with its default configuration values (as provided in WEKA) using command line interface (CLI).

Table I also lists information about an algorithm called DecisionStump. Classification accuracy of simple learning algorithms (weak learners), e.g., decision trees, naïve bayes, can be affected by potential bias in the training dataset. Thus, ensemble classifiers are used to improve their classification accuracy. AdaBoost1 belongs to a class of ensemble classifiers (boosting) that help to improve the classification accuracy of weak learners by training them iteratively, with different sets of weights assigned to class labels in each iteration. WEKA's default configuration of *AdaBoost1* implements a meta classifier that improves the accuracy of the model built using *DecisionStump,* a tree-based classifier (weak learner).

Algorithm	Subject Programs	Number of Branches	Number of Mutants
j48	weka.classifiers.trees.j48*	750	3796
NaiveBayes	weka.classifiers.bayes.NaiveBayes.java	203	1075
AdaBoost1	weka.classifiers.meta.AdaBoost1.java	90	491
DecisionStump	weka.classifiers.trees.DecisionStump.java	128	921
OneR	weka.classifiers.rules.OneR.java	88	510

Table I – Information about subject programs

C. Datasets

We identify suitable datasets from Kaggle.com, which provides access to public databases. By default, dataset search results on Kaggle.com are sorted by hotness, a measure indicative of the amount of interests and recency of datasets on their platform [9]. Other methods of sorting include *New, Recently Active, Most Votes, Updated and Relevance.* As Kaggle.com does not release the hotness calculation formula to the public [10], we are not completely clear of how the hotness of datasets is computed. Hence, we sort the search results by *Most Votes*, which sorts datasets based on the most popular datasets of all time. Then, the results are further filtered with the following two criteria: (a) *size – 10 MB to 1GB* and (b) *File types – CSV.* Next, we inspect each dataset in the order sorted by Kaggle.com and select datasets that require no cleaning and can be executed in WEKA.

We identified five datasets from different application domains, including *AustralianWeather [23], ForestCover [24, 25], Crime [26]*, *SupplyChain* [21] and *VideoGames [22].* The *ForestCover* dataset is a multi-label classification dataset with seven different class labels. The remaining four datasets consist of binary class labels. Table II lists the datasets and their information.

We selected datasets such that data preprocessing is minimal. No modification was required for *AustralianWeather and SupplyChain* as their respective class labels were nominal by default. The class labels of the remaining three datasets, i.e. *ForestCover, Crime and VideoGames* were converted from numeric to nominal using WEKA's built-in filter.

DATASET	# OF CLASS LABELS	# OF INSTANCES	# OF ATTRIBUTES
ForestCover	7	581,012	55
AustralianWeather	2	142,193	23
Crime	2	284,807	31
SupplyChain	2	580,251	5
VideoGames	2	999,999	56

TABLE II – DATASET INFORMATION

D. Generation of Reduced Datasets

For each original dataset in Table II, two groups of smaller datasets are generated. Group 1 consists of reduced datasets generated using pure random sampling, whereas in Group 2, reduced datasets are generated using stratified sampling. Recall that stratified sampling maintains the overall class distribution of the original datasets. For each group, we generate samples of four different sizes, i.e., 100, 200, 400, 800. Also, in order to reduce variations in random sampling, we generate five samples for each sample size by using different random seeds. Thus, each dataset has 20 samples per group and a total of 40 samples in the two groups.

WEKA provides a set of pre-processing filters that allow users to modify datasets. Reduced datasets in Group 1 (random sampling) are generated using WEKA's pre-processing filter *weka.filters.**unsupervised**.instances. Resample*. Reduced datasets in Group 2 (stratified sampling) are generated using pre-processing filter *weka.filters.**supervised**.instances.Resample*. These filters allow the user to select the sample size, usually specified as a percentage of the original dataset. Note that both filters perform a volumetric reduction, i.e. the number of instances in the dataset is reduced whereas the number of attributes will remain unchanged.

For example, consider a dataset of 100,000 data instances with four class labels, A, B, C and D. Assume that their class distribution is as follows: 30% instances belong to Class A, 40% instances belong to Class B, 10% instances belong to Class C and the remaining 20% belongs to Class D. Generating a smaller dataset with 100 instances using stratified sampling (Group II) will consists of 30 instances belonging to Class A, 40 instances belonging to Class B, 10 instances belonging to Class C and 20 instances belonging to Class D. In contrast, samples generated using random sampling (Group I) does not necessarily maintain the class label distribution.

The *Crime dataset* (284,807 instances) has the following class distribution: 99.82% instances belong to Class 0 (284,315 instances), and 0.18% instances belong to Class 1 (492 instances). When generating a reduced dataset with 800 instances using WEKA's pre-processing filter, it is highly likely that random sampling fails to produce a reduced dataset that include instances in both Class 0 and Class 1. Instead, it is likely that all of the 800 instances belong to Class 0. A developer might face the above said scenario when s/he generates a reduced dataset using random sampling from a class-imbalanced (or skewed) dataset. As a workaround, a developer can create a reduced dataset while preserving the original class distribution. This is our motivation to use two different groups of samples and to investigate their impact in testing supervised learning algorithms. The original datasets and their reduced versions are made publicly available at [32].

E. Metrics

We use both branch coverage and mutation coverage to measure test effectiveness. Branch coverage is recorded using JaCoCo [18]. Mutation coverage is obtained using PITest (PIT), which is a widely used mutation testing framework [19]. PIT can automatically seed one fault at a time into SUT and execute the mutated code against the unit test(s) specified. In our experiments, we have thirteen mutation operators including all the default mutators (seven), three experimental mutators and three optional mutators [20, 31].

Given the nature of our subject algorithms, there may be loops that could be executed for a large number of times. In this scenario, it is possible that PITest mistakenly concludes that an algorithm has entered an infinite loop and thus kills the execution, thus resulting in a false positive. To prevent this scenario, we set the timeout constant factor to 100000 milliseconds. The reason for choosing 100000 milliseconds is based on our observation that among 20 baseline executions (5 original datasets x 4 algorithms), 16 executions (80%) completed under 50 seconds. Therefore, we choose the timeout constant to be 2 x 50 = 100 seconds.

The machine we used for our experiments is a workstation with two Xeon E5- 2630V3 8 core CPUs @ 2.40GHz, 64GB DDR4 2133 MT/s memory, and a Samsung 850 EVO 500GB SSD.

III. EXPERIMENTAL RESULTS

In this section, we present our experimental results and discussion about our results. In Section III.A, we present the branch coverage results achieved by the original datasets. These results are considered to be the baseline results. In Section III.B, we present the branch coverage results achieved by the reduced datasets. These results are compared to the baseline results. In Section III.C, we present the mutation coverage results achieved by both of the original and reduced datasets.

A. Branch Coverage of the Original Datasets

Table III presents the branch coverage achieved by algorithms with original datasets. Among the datasets, *SupplyChain* consistently achieve higher coverage for all the algorithms. We observe that across algorithms, a considerable number of methods, and their branches were not executed, and thus the overall branch coverage appears to be considerably lower (<= 50%). This, however, can be explained as follows. Consider the branch coverage results of the OneR algorithm. The *SupplyChain* dataset achieves the highest branch coverage (57%), i.e., 51 out of 88 total branches. Among the missing 37 branches, 18 branches missed due to default configuration options. Seven branches are related to error handling, such as missing attribute values, and the remaining 12 branches cannot be covered as cross-validation is not performed while building models using the command-line interface (CLI).

To our surprise, *AustralianWeather* covers a significantly smaller number of branches (17) compared to the rest. This can be explained as follows: Among the five datasets, all the attributes of *AustralianWeather* belong to the nominal data type. All the attributes of *ForestCover, VideoGames, and Crime* belong to the numeric data type. In the case of *SupplyChain,* 3 out of 4 attributes belong to the numeric data type, and the remaining attribute belongs to the nominal data type. When executing the OneR algorithm with *AustralianWeather,* a method, newNumericRule(), was missed that has 36 branches and handles numeric attributes. Hence, *AustralianWeather* achieves a significantly lower branch coverage, whereas *SupplyChain* achieves the highest branch coverage, as it covers branches related to both numeric and nominal data types.

In our experiments, we executed the algorithms using WEKA's default configuration options only. This could cause branching conditions that are specific for other configuration options to be missed. As shown in [2], executing different configuration options could significantly increase branch coverage. Also, the branches related to error handling and GUI are not covered as we run our tests with clean datasets using the CLI.

We emphasize that, although branch coverage achieved by original datasets is not high, this does not affect the purpose of our experiments, which is to determine whether reduced datasets could achieve the same or similar coverage as the original dataset.

Datasets	Algorithms	# of Branches Covered	Total Number of Branches	Branch Coverage
AustralianWeather	j48	180	750	24%
ForestCover		202		26%
SupplyChain		201		26%
VideoGames		202		26%
Crime		195		26%
AustralianWeather	Naïve Bayes	73	203	35%
ForestCover		77		37%
SupplyChain		99		48%
VideoGames		79		38%
Crime		78		38%
AustralianWeather	AdaBoost1	28	90	31%
ForestCover		17		18%
SupplyChain		28		31%
VideoGames		28		31%
Crime		28		31%
AustralianWeather	DecisionStump	50	128	39%
ForestCover		47		36%
SupplyChain		71		55%
VideoGames		48		37%
Crime		48		37%
AustralianWeather	OneR	17	88	19%
ForestCover		44		50%
SupplyChain		51		57%
VideoGames		45		51%
Crime		45		51%

TABLE III - BRANCH COVERAGE FOR ORIGINAL DATASETS

B. Branch Coverage of Reduced Datasets

In this section, we present the branch coverage results achieved by reduced datasets. For each dataset, we generate reduced datasets using two different approaches: random sampling and stratified sampling; we generate reduced datasets in four different sizes: 100 instances, 200 instances, 400 instances, and 800 instances, as discussed in Section II-D. Due to limited space, we present the median branch coverage achieved by each size relative to their baseline coverage.

Tables IV and V present the branch coverage results of reduced datasets generated using random sampling and stratified sampling, respectively. All the coverage results presented here are relative to their corresponding baseline. i.e., a relative branch coverage of 1.0 suggests that a reduced dataset achieves a branch coverage identical to the original dataset. Note that, in Tables IV and V, 39 out of 50 reduced datasets of size 800 produced by both random and stratified sampling, achieved branch coverages identical to the baseline; for the remainder of the cases, we notice the coverages do not significantly vary among different sample sizes. Therefore, in our experiments we did not consider sample size larger than 800 instances.

The results indicate that, for the j48 algorithm, reduced datasets of size 800 instances produced by both random and stratified sampling of *ForestCover, SupplyChain, and VideoGames* can retain their baseline branch coverage. For the NaiveBayes algorithm, the reduced versions of all five datasets can retain their branch coverage achieved by their respective original datasets and in some cases, reduced datasets achieving even higher branch coverage. Similarly, for the remaining three algorithms namely AdaBoost1, DecisionStump, and OneR, the reduced versions of all datasets except *Crime*, in most cases either retain their respective baseline branch coverage (1.0) or in some cases achieve a branch coverage closer to its baseline (0.9<=branch coverage<1.0).

For the reduced datasets of *Crime,* we observe that three out of five algorithms (j48, AdaBoost1, One-R) suffer from a loss in branch coverage. In particular, consider the case of j48 (Row 5 in Tables IV and V), which suffers from a significant loss in branch coverage. This is attributed to the class imbalance problem. The *Crime* dataset consists of 284,807 instances with two class labels: (0, 1); 99.82% instances belonging to Class 0 and remaining 0.18% belonging to Class 1. Due to class imbalance, chances of drawing all hundred samples (at random) that belong to Class 0 is higher.

In our experiments, for the reduced datasets of size 100 produced by random sampling, four out of five samples have all their instances belonging to Class 0, and they achieve a relative median branch coverage of 0.12. On the contrary, three out of five reduced datasets of size 200 produced by random sampling have representation from both of the class labels, and they achieve a higher branch coverage comparatively (0.35). We notice that, in the case of j48, if a reduced dataset consists of a single label, there is a significant loss in branch coverage.

Next, we compare the coverage results of random sampling and stratified sampling. Our results indicate that, in most cases, the datasets reduced using both random and stratified sampling can achieve the same branch coverage.

Datasets	Algorithms	Size of the reduced dataset			
		100	200	400	800
AustralianWeather	j48	0.75	0.75	0.71	0.71
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		0.81	0.73	0.92	1.00
VideoGames		0.96	0.96	0.96	1.00
Crime		0.12	0.35	0.12	0.35
AustralianWeather	Naïve Bayes	1.00	1.00	1.00	1.00
ForestCover		1.03	1.03	1.03	1.03
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		1.00	1.00	1.00	1.00
AustralianWeather	AdaBoost1	1.00	1.00	1.00	1.00
ForestCover		1.78	1.67	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.65	0.65	0.65	0.65
AustralianWeather	DecisionStump	0.95	0.95	0.95	0.95
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.97	1.00	0.97	1.00
AustralianWeather	OneR	0.95	0.95	0.95	0.95
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		0.96	0.96	0.96	0.96
VideoGames		1.00	1.00	1.00	1.00
Crime		0.76	0.92	0.76	1.00

Table IV – Relative Branch Coverage of reduced datasets (Random Sampling)

Datasets	Algorithms	Size of the reduced dataset			
		100	200	400	800
AustralianWeather	j48	0.75	0.75	0.71	0.71
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		0.81	0.92	0.96	1.00
VideoGames		0.92	0.96	1.00	1.00
Crime		0.12	0.12	0.12	0.35
AustralianWeather	Naïve Bayes	1.00	1.00	1.00	1.00
ForestCover		1.03	1.03	1.03	1.03
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		1.00	1.00	1.00	1.00
AustralianWeather	AdaBoost1	1.00	1.00	1.00	1.00
ForestCover		1.78	1.67	1.78	1.67
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.65	0.65	0.65	1.00
AustralianWeather	DecisionStump	1.00	1.00	1.00	1.00
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.97	0.97	0.97	1.00
AustralianWeather	OneR	0.95	0.95	0.95	0.95
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		0.96	0.96	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.76	0.76	0.76	0.92

Table V – Relative Branch Coverage of Reduced Datasets (Stratified Sampling)

In the cases of *AustralianWeather*, *SupplyChain* and *VideoGames*, the datasets reduced using both random and stratified sampling achieves identical branch coverage. This can be explained by the fact that all reduced datasets have a good class label representation. For example, all five sample datasets of *AustralianWeather* of size 100 that are reduced using stratified sampling have the following class label distribution: 78 instances belong to *No*, and 22 instances belong to *Yes.* In the case of random sampling, amongst five samples, sample 5 consists of 86 instances belong to *No* and 14 instances belongs to *Yes* whereas, Sample 3 consists of 74 instances belong to *No* and 26 instances belongs to *Yes.*

Our results indicate that the reduced datasets of *ForestCover* generated using both random and stratified sampling achieve the same branch coverage as the original datasets across all algorithms. In comparison, the reduced datasets generated from *AustralianWeather*, *SupplyChain*, *VideoGames*, and *Crime* suffer from a minimal to moderate coverage loss in at least one of the five algorithms. This may be attributed to the fact, *ForestCover* is a multilabel dataset (7 class labels), whereas the rest of the four datasets are binary label dataset. More experimental data is required to obtain a better understanding. Also, our results indicate that in the case of the AdaBoost1 algorithm, the reduced datasets achieve a better branch coverage compared to the baseline, i.e., the original datasets. To some extent, this result is surprising, given the significant increase in branch coverage. This is possible because the reduced datasets may trigger execution scenarios that are different than the original datasets.

In the case of the *Crime* dataset, three algorithms suffer from a coverage loss. In particular, consider the coverage achieved by the reduced datasets of *Crime* produced by both random and stratified sampling. Row 5 in Tables IV and V indicates that the reduced dataset of size 200 produced by random sampling achieves a higher branch coverage (0.35) compared to the reduced dataset produced by stratified sampling of the same size (0.12). This can be attributed to the representativeness of the class label. On examination of reduced datasets, we observe that three out of five samples generated using random sampling have instances belonging to two class labels (Class 0 and Class 1). However, in the case of datasets reduced using stratified sampling, all instances belong to a single class (Class 0). Hence, subject programs achieve lower coverage while executing with stratified samples as they fail to trigger the execution of certain branches. The branch coverage results of the OneR algorithm suggest a similar pattern, i.e., the reduced dataset of size 200 produced by random sampling achieves a higher coverage (0.92) compared dataset reduced using stratified sampling of the same size (0.76).

This behavior of stratified sampling, i.e., all the instances of a reduced dataset belonging to a single class, is expected as it draws samples in a way that maintains the class distribution of the original dataset. Recall that the *Crime* dataset consists of 284,807 instances with two class labels: (0, 1); 99.82% instances belonging to Class 0 and remaining 0.18% belonging to Class 1. To generate a reduced dataset of size 200 instances using stratified sampling, instances are drawn in the following way (99.82% * 200) > 199 (instances) belonging to Class 0 and (0.18% * 200) < 1 (instances) belonging to Class 1. Hence, all the instances belong to Class 0 and thus, the reduced dataset suffers from lack of class representativeness.

For the *Crime* dataset, a minimum of 556 instances is required to guarantee that a reduced dataset (stratified sampling) consists of instances belonging to both classes (0 and 1). Among four different sizes (100, 200, 400, and 800) of reduced datasets generated using stratified sampling, in three groups (100,200 and 400), all instances belong to class 0 and thus achieve a low branch coverage (0.12). In the case of reduced datasets of 800 instances, all five samples consist of instances of both classes and thus achieve a relatively higher branch coverage (0.35).

Our results indicate that approximately 80% of the reduced datasets achieve coverage identical or similar to the original datasets. In another word, the volume of a dataset does not directly attribute to branch coverage. Instead, factors such as lack of representativeness of class labels in a reduced dataset could impact branch coverage. The results suggest that in most cases, reduced datasets do not suffer from branch coverage loss. In this respect, they can be used in place of the original datasets to speed up the testing process.

Among the two sampling approaches, the results indicate that in most cases (around 75%) reduced datasets generated using both random and stratified sampling exhibit identical behavior. However, when a tester decides to use stratified sampling, he/she should choose the size of the reduced dataset (minimum number of samples) based on the original class distribution such that each class label is represented in the reduced dataset.

C. *Mutation Coverage of Reduced Datasets*

In this section, we present the mutation coverage results achieved by algorithms while executing with reduced datasets.

Given the size of the datasets and the number of mutants generated for SUT, the overall execution time can be between a few hours to several days. Due to time constraints, our experiments have an execution time limit of 48 hours (chosen arbitrarily). If a dataset takes more than 48 hours to complete, then we kill the test execution and use a relatively smaller dataset (10000 instances) as our baseline. Out of 20 baseline test executions, one baseline execution, j48 algorithm with the *VideoGames* dataset executed for more than 2 days. Hence, we generated five smaller samples of *VideoGames* dataset with 10000 instances each and used their median coverage as a baseline.

DATASETS	ALGORITHMS	SIZE OF THE REDUCED DATASET			
		100	200	400	800
AustralianWeather	j48	0.50	0.50	0.44	0.50
ForestCover		0.96	0.96	0.96	1.00
SupplyChain		0.64	0.57	0.71	0.79
VideoGames		0.88	0.88	0.92	0.96
Crime		0.14	0.24	0.14	0.24
AustralianWeather	Naïve Bayes	0.94	0.94	0.94	0.94
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		1.00	1.00	1.00	1.00
AustralianWeather	AdaBoost1	1.00	1.00	1.00	1.00
ForestCover		1.92	1.38	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.04	1.00	1.00	1.00
Crime		0.50	0.54	0.50	0.54
AustralianWeather	DecisionStump	1.00	1.00	1.00	1.00
ForestCover		1.03	1.00	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.85	0.94	0.85	0.94
AustralianWeather	OneR	0.93	0.93	0.93	0.93
ForestCover		0.97	1.00	1.00	1.00
SupplyChain		1.07	1.07	1.07	1.07
VideoGames		0.97	0.97	1.00	1.00
Crime		0.66	0.77	0.66	0.89

TABLE VI - RELATIVE MUTATION COVERAGE OF REDUCED DATASETS (RANDOM SAMPLING)

DATASETS	ALGORITHMS	SIZE OF THE REDUCED DATASET			
		100	200	400	800
AustralianWeather	j48	0.50	0.50	0.50	0.50
ForestCover		0.92	0.96	1.00	0.96
SupplyChain		0.64	0.71	0.79	0.79
VideoGames		0.54	0.88	0.96	0.96
Crime		0.14	0.14	0.14	0.24
AustralianWeather	Naïve Bayes	0.94	0.94	0.94	0.94
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		1.00	1.00	1.00	1.00
AustralianWeather	AdaBoost1	1.00	1.00	1.00	1.00
ForestCover		2.00	1.38	2.00	1.38
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.50	0.50	0.50	1.00
AustralianWeather	DecisionStump	1.00	1.00	1.00	1.00
ForestCover		1.03	1.00	1.03	1.00
SupplyChain		1.00	1.00	1.00	1.00
VideoGames		1.00	1.00	1.00	1.00
Crime		0.85	0.85	0.85	0.97
AustralianWeather	OneR	0.93	0.93	0.93	0.93
ForestCover		1.00	1.00	1.00	1.00
SupplyChain		1.07	1.07	1.13	1.07
VideoGames		0.97	1.00	0.97	1.00
Crime		0.66	0.66	0.66	0.77

TABLE VII-RELATIVE MUTATION COVERAGE OF REDUCED DATASETS (STRATIFIED SAMPLING)

Tables VI and VII present the mutation coverage results of the reduced datasets. All the coverage results presented here are relative to their corresponding baseline. The results from Tables VI and VII suggest that the j48 algorithm performs poorly with the reduced datasets of *AustralianWeather* and *SupplyChain.* Similarly, the reduced datasets of *Crime* result in a mutation coverage decrease for all the algorithms except Naive Bayes. The rest of the reduced datasets generated using both random and stratified sampling can retain their baseline mutation coverage.

We report that the majority of the mutation coverage results (except reduced datasets of *SupplyChain* on j48) mirrors with their respective branch coverage results (Table IV & V; Table VI & VII).

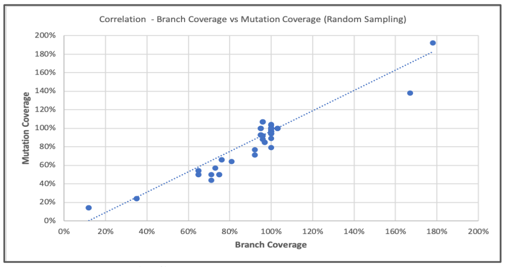


FIGURE 1 – CORRELATION GRAPH – RANDOM SAMPLING

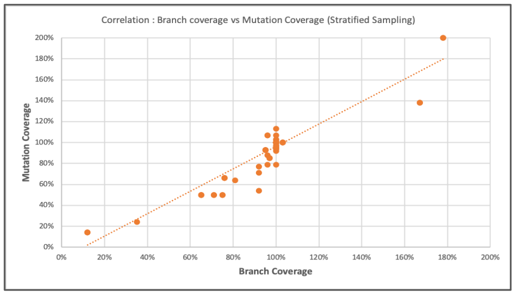


FIGURE 2 – CORRELATION GRAPH – STRATIFIED SAMPLING

Figures 1 and 2 present a correlation graph of branch coverage vs. mutation coverage for random sampling and stratified sampling, respectively. In Figures 1 and 2, x-axis indicates branch coverage, and the y-axis indicates mutation coverage. For the datasets reduced via random sampling, branch vs. mutation coverage has a Pearson correlation coefficient of 0.944148, whereas the datasets reduced via stratified sampling has a fractionally lower Pearson correlation coefficient of 0.939506. The result suggests that in most cases, mutation coverage has a strong positive correlation with the branch coverage. To our surprise, the mutation results of j48 using the *SupplyChain* dataset reduced using stratified sampling does not appear to correlate well with branch coverage, and we plan to investigate this further as part of our future work.

IV. THREATS TO VALIDITY

Threats to internal validity are factors that may be responsible for experimental results, without our knowledge. To reduce human errors in the experimental procedure, we tried to automate our experiments as much as possible. In particular, we wrote scripts to automatically execute tests, measure code and mutation coverage, and generate coverage reports. Further, the results generated from samples of each dataset were verified manually, whenever possible.

Threats to external validity occur when the experimental results could not be generalized to other subjects. Using a single dataset for our experiments might impact the validity of our results due to lack of representativeness. To mitigate this threat, we used four supervised learning algorithms from WEKA that belong to different groups and five datasets from different application domains. More experiments using other learning algorithms, including both supervised and unsupervised algorithms, and other datasets, can further reduce the threats to external validity.

V. RELATED WORK

First, we review existing work reported on testing machine learning algorithms. One challenge in testing machine learning algorithms is how to deal with the test oracle problem. Murphy et al. [4,5] proposed a metamorphic testing technique to test machine learning algorithms. They developed metamorphic properties for three machine learning algorithms, including MartiRank, SVMLight, and PAYL. Similarly, Nakajima et al. [7] proposed a systematic approach to derive metamorphic properties and translation functions for testing a special class of classifiers known as Support Vector Machines (SVM). Xie et al. [11] proposed a metamorphic testing approach to test supervised learning algorithms, namely Naïve Bayes classifier and k-nearest neighbor classifier. Our work differs from these works in that we focus on evaluating the effectiveness of using smaller datasets in testing supervised learning algorithms.

Second, we review existing work on dataset reduction for testing big data applications. Such work is relevant because many machine learning algorithms are big data applications in that they are designed to learn from large amounts of data. Ur Rehman et al. [13] reviewed existing data reduction techniques such as compression-based data reduction method, dimension reduction techniques for big data applications. Czarnowski et al. [14] proposed an agent-based population learning algorithm for data reduction. Their algorithm aims at finding a subset of the original dataset that can be used to build a classifier that is similar to the classifier built using the original dataset. This is different from our work, which tries to find a subset of the original dataset that preserves test effectiveness.

We mention that a significant amount of work is reported on data reduction techniques in terms of dimensionality reduction and feature space [3, 6, 8, 16, 17]. In contrast, our work focuses on volume reduction, i.e., reducing the number of instances in a big dataset. To the best of our knowledge, our work is the first to investigate the effectiveness of volume reduction in testing machine learning algorithms.

VI. CONCLUSION AND FUTURE WORK

In this paper, we report a study that investigates the use of reduced datasets in testing machine learning algorithms. We used four supervised learning algorithms from WEKA as our subject programs. Five publicly available datasets from Kaggle.com were chosen as subject datasets. For each dataset, we generated reduced datasets in four different sizes using random and stratified sampling. Then, we executed the algorithms with the original and the reduced datasets and measured test effectiveness in terms of branch and mutation coverage. Our results indicate, in most cases, reduced datasets of very small sizes (e.g. 800 instances) can retain branch and mutation coverage of the original, big datasets (e.g., >100,000 instances). This suggests that reduced datasets can be used to effectively test machine learning algorithms. Our results also indicate a high correlation between branch coverage and mutation coverage. Thus, branch coverage can be used when mutation testing is prohibitively expensive.

This is the first step in our larger effort to speed up testing machine learning algorithms. We plan to continue our work in the following directions. First, we plan to investigate the reduction of even bigger multi-label datasets (> 1 GB) and its effect on testing machine learning algorithms. Second, we plan to expand our study to include unsupervised learning algorithms. Compared to supervised learning algorithms, unsupervised learning algorithms learn from unlabeled datasets and thus could be harder to validate its output. Third, our experiments show that there exists a high correlation between branch and mutation coverage. However, some recent work reports that traditional code coverage measures such as branch coverage may not be adequate for testing deep learning algorithms. We believe that this has to do with the nature of the algorithms and also the types of fault that may exist in the algorithms. We plan to study this further by conducting experiments on deep learning algorithms. Finally, we plan to develop new methods, i.e., methods other than random sampling, for dataset reduction. For example, we are investigation how to perform equivalence partitioning among instances in a big dataset, and then choose one or more representatives from each equivalence group.

VII. Acknowledgement

This work is supported by research grant (70NANB15H199) from Information Technology Lab of National Institute of Standards and Technology (NIST).

Disclaimer: Certain software products are identified in this document. Such identification does not imply recommendation by the NIST, nor does it imply that the products identified are necessarily the best available for the purpose.

References

[1] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1.

[2] Chandrasekaran, Jaganmohan, et al. "Applying combinatorial testing to data mining algorithms." *2017 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. IEEE, 2017.

[3] Feldman, Dan, Melanie Schmidt, and Christian Sohler. "Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering." *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2013.

[4] Murphy, Christian, Gail E. Kaiser, and Marta Arias. "An approach to software testing of machine learning applications." (2007).

[5] Murphy, Christian, Gail E. Kaiser, and Lifeng Hu. "Properties of machine learning applications for use in metamorphic testing." (2008).

[6] Kira, Kenji, and Larry A. Rendell. "The feature selection problem: Traditional methods and a new algorithm." *Aaai*. Vol. 2. 1992.

[7] Nakajima, Shin, and Hai Ngoc Bui. "Dataset coverage for testing machine learning computer programs." *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2016.

[8] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." *2014 Science and Information Conference*. IEEE, 2014.

[9] https://www.kaggle.com/docs/datasets.

[10] https://www.kaggle.com/general/39290

[11] Xie, Xiaoyuan, et al. "Testing and validating machine learning classifiers by metamorphic testing." *Journal of Systems and Software* 84.4 (2011): 544-558.

[12] Zhang, Zhiyi, and Xiaoyuan Xie. "Towards testing big data analytics software: the essential role of metamorphic testing." *Biophysical reviews* 11.1 (2019): 123-125.

[13] ur Rehman, Muhammad Habib, et al. "Big data reduction methods: a survey." *Data Science and Engineering* 1.4 (2016): 265-284.

[14] Czarnowski, Ireneusz, and Piotr Jędrzejowicz. "An Approach to Data Reduction for Learning from Big Datasets: Integrating Stacking, Rotation, and Agent Population Learning Techniques." *Complexity* 2018 (2018).

[15] Czarnowski, Ireneusz, and Piotr Jędrzejowicz. "Stacking and rotation-based technique for machine learning classification with data reduction." *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*. IEEE, 2017.

[16] Liu, Qingzhong, et al. "Mining the big data: The critical feature dimension problem." *2014 IIAI 3rd International Conference on Advanced Applied Informatics*. IEEE, 2014.

[17] Wold, Svante, Kim Esbensen, and Paul Geladi. "Principal component analysis." *Chemometrics and intelligent laboratory systems* 2.1-3 (1987): 37-52.

[18] M. Hoffmann, B. Janiczak, E. Mandrikov and M. Friedenhagen. Jacoco code coverage tool. Online , 2016

[19] H. Coles. Pit mutation testing. http: //pitest.org/, 2016.

[20] Coles, Henry, et al. "Pit: a practical mutation testing tool for java." *Proceedings of the 25th International Symposium on Software Testing and Analysis*. ACM, 2016.

[21] https://www.kaggle.com/rtatman/lego-database#inventory_parts.csv

[22] https://www.kaggle.com/paololol/league-of-legends-ranked-matches#stats1.csv

[23] https://www.kaggle.com/jsphyg/weather-dataset-rattle-package

[24] As Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science

[25] https://www.kaggle.com/c/forest-cover-type-prediction/overview

[26] https://www.kaggle.com/mlg-ulb/creditcardfraud

[27] John, George H., and Pat Langley. "Estimating continuous distributions in Bayesian classifiers." *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1995.

[28] Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." *icml*. Vol. 96. 1996.

[29] Holte, Robert C. "Very simple classification rules perform well on most commonly used datasets." *Machine learning* 11.1 (1993): 63-90.

[30] Salzberg, Steven L. "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993." *Machine Learning* 16.3 (1994): 235-240.

[31] https://pitest.org/quickstart/mutators/

[32] https://1drv.ms/u/s!AjZ3W-Mz9wPKhtlLoWUU2zZKzm4bRg?e=QT8Oko

AC and DC Quantized Hall Array Resistance Standards

Randolph E. Elmquist[*], Mattias Kruskopf[*,§] , Dinesh K. Patel[*,†], I-Fan Hu[*,†], Chieh-I Liu[*,‡], Albert F. Rigosi[*], Alireza R. Panna[*], Shamith U. Payagala[*], and Dean G. Jarrett[*]

[*]National Institute of Standards and Technology, 100 Bureau Drive, Stop 8171, Gaithersburg, MD, 20899, USA
elmquist@nist.gov

[§]Physikalisch-Technische Bundesanstalt, Department of Electrical Quantum Metrology, Braunschweig, 38116, Germany

[†]Graduate Institute of Applied Physics, National Taiwan University, Taipei 10617, Taiwan

[‡]University of Maryland, Department of Chemistry and Biochemistry, College Park, MD, 20742, USA

Abstract — **Quantized Hall array resistance standards (QHARS) span values from 100 Ω to 1 MΩ and demonstrate precision approaching that of single devices. This paper focuses on QHARS having values near 1 kΩ for increased sensitivity using room-temperature direct current comparator (DCC) bridges and digital impedance bridges. We introduce a dc QHARS design that uses the Wheatstone bridge principle as a precise test of the QHARS self-consistency for dc measurements. Branched contacts and superconducting interconnections help to reduce ac and dc loading and leakage errors with near-zero resistance at magnetic flux densities of 9 T.**

Index Terms **— electrical measurement standards, quantized Hall resistance, epitaxial graphene, Wheatstone bridge.**

I. Introduction

Equivalent-circuit models of multi-terminal devices [1] have been employed to explore loading and contact resistance effects in measurements of R_H, the quantized Hall resistance (QHR). The main observation is that current drawn from the Hall voltage terminals can cause significant loading error due to the effective series source resistance $r_s = R_H/2$ between the contact (reservoir) and the edge state of the QHR device [2] in a strong magnetic field. In 1993, the metrological application of these principles was established by designs of circuits with multiple links between two or more devices [3]. The first link carries the majority of the current and sets up the equipotential edges on each device, so that Hall voltage interconnections have much smaller loading currents. Thus, loading and dc contact resistance effects can be reduced to negligible levels in QHARS networks. Likewise, multiple connections minimize the effects of parasitic loading in impedance measurements of single devices, and the development of QHR standards in the audio-frequency range has been based on this advance.

II. Ac measurement standards

Prior work in the metrology community with ac QHR standards [4], [5] has not incorporated QHARS devices, since complex circuit interconnections increase the parasitic capacitance. Positive contributions to the frequency-dependent impedance [4] may occur when capacitive current flows between the edge states and conductors that shield the device. Loss occurs only at the edges because the Corbino effect prevents dissipation from the interior, incompressible state of the two-dimensional electron gas. Similarly, ac current across the capacitive junction between the two equipotential edges creates a negative impedance contribution. We have recently reported on epitaxial graphene QHE devices with single-layer NbTiN superconducting interconnections [6], [7] and propose to use this technique for developing ac QHARS having 13 devices in parallel, yielding a resistance near 992.8 Ω. Confocal optical microscope images of sections of prototype arrays are shown in Fig. 1. Superconducting connections include branching contacts to the source and drain [7] and do not require isolated crossovers of the current and potential leads.

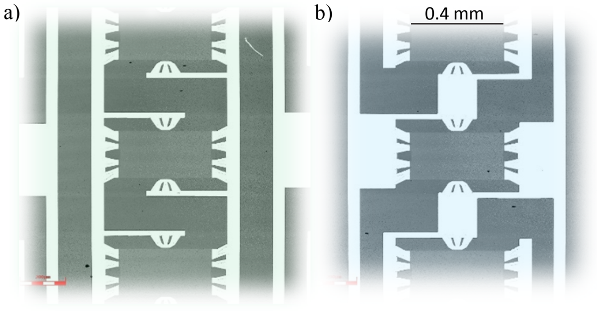


Fig. 1. QHARS device arrays. (a) Current flows in the same direction for all devices. (b) Current flows in alternating directions for adjacent devices to reduce capacitive losses between devices.

III. Dc measurement standards

In comparison with cryogenic current comparator (CCC) systems, room-temperature DCC bridges are less expensive and can be controlled through intuitive automated interfaces. Resistance standards at 1000 Ω are essential for DCC metrology based on the $\nu = 2$ plateau ($R_H \approx 12.9064$ kΩ) and for DCC scaling measurements between 100 Ω and 10 kΩ.

Elmquist, Randolph; Kruskopf, Mattias; Patel, Dinesh; Hu, I Fan; Liu, Chieh-I; Rigosi, Albert; Panna, Alireza; Payagala, Shamith; Jarrett, Dean G. "AC and DC Quantized Hall Array Resistance Standards." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020.

We previously reported on scaling results for DCC measurements based on a single device [7], finding Type A uncertainty about five times smaller for ratios of 1 kΩ/100 Ω than for 12.9 kΩ/1 kΩ with the same current of 1 mA applied to the 1 kΩ standard. Lower uncertainty results for a 1 kΩ/100 Ω ratio because the 1 kΩ measurement current flows in the more sensitive primary winding. Laboratories using DCC bridges could thus improve SI traceability by using QHARS devices at 1 kΩ as the realization of the unit.

Designs of some QHARS incorporate hundreds of devices at present [8] – [11], and superconducting connection designs may allow even higher density integration [6]. The concern that we address for wider adoption of QHARS technology at dc is the reliability of quantum SI traceability. Recommended longitudinal resistance and contact resistance measurements, used for single QHR devices, are impractical at large scales.

Figure 2 shows a dc QHARS circuit with superconducting interconnections. Two identical series-parallel (S/P) arms formed by sets of 11 and 15 QHR devices create a balanced Wheatstone bridge at two null points with voltage V_{SD} applied at the source and drain. A resistance of ≈1016.868 Ω is realized by the series-parallel arms. Three sets of 14 devices in series (S) between the source and drain trim the resistance correction to ≈ -14.696 μΩ/Ω from the nominal 1 kΩ value, and also have intermediate terminals. The design of the array allows an estimation of QHARS resistance error from the consistency of null-voltage measurements at the intermediate terminals with voltages of $V_{\mathrm{S/P}} \approx 0.577\ V_{\mathrm{SD}}$ and $V_{\mathrm{S}} = 0.5\ V_{\mathrm{SD}}$.

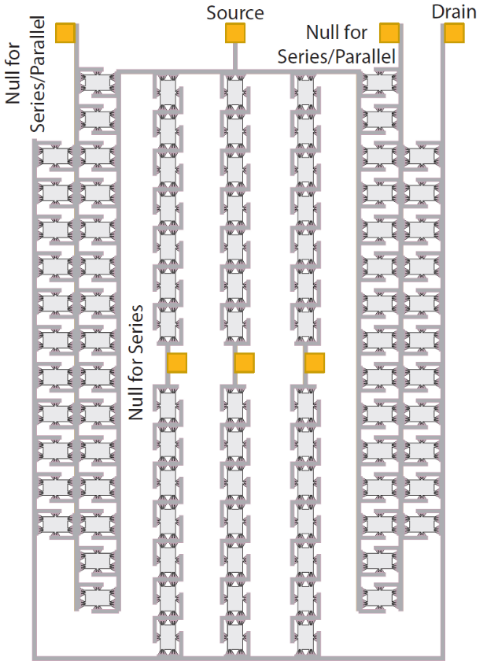


Fig. 2. QHARS array of 94 QHR devices with symmetric Wheatstone bridge elements allows null voltage measurements between intermediate contacts at similar voltage.

IV. Conclusion

Our approach condenses the multi-series interconnection principle by using superconducting NbTiN junctions to create two-terminal series and parallel quantum Hall arrays, thus eliminating possible errors from leakage and reducing internal loading errors in QHARS circuits. The capacitive parasitic loading error is minimized, since loading current is supplied by the source and drain, and this may allow parallel QHARS arrays to provide an alternative to single devices for use in ac bridges. To improve dc reliability, we will apply the Wheatstone bridge principle for *in-situ* characterization of QHARS circuits so that any contact resistance and longitudinal resistance contributions also may be detected for reduced uncertainty.

Acknowledgement

The work of D. K. Patel and I.-F. Hu at NIST was made possible by arrangement with Prof. C.-T. Liang of National Taiwan University.

References

[1] B. W. Ricketts and P. C. Kemeny, “Quantum Hall effect devices as circuit elements,” *J. Phys. D: Appl. Phys*. vol. 21, p. 483, 1988.

[2] M. Büttiker, “Absence of backscattering in the quantum Hall effect in multiprobe conductors,” *Phys. Rev. B*, vol. 38, p. 9375, 1988.

[3] F. Delahaye, “Series and parallel connection of multiterminal quantum Hall-effect devices,” *J. Appl. Phys.* vol. 73, p. 7914, 1993.

[4] J. Schurr, F. Ahlers, and B. P. Kibble, “The ac quantum Hall resistance as an electrical impedance standard and its role in the SI,” *Meas. Sci. Technol.* vol. 23, no. 12, p. 124009, 2012.

[5] C.-C. Kalmbach, *et al.*, “Towards a graphene-based quantum impedance standard, *Appl. Phys. Lett.* vol. 105, p. 073511, 2014.

[6] M. Kruskopf *et al.*, “Next-generation crossover-free quantum Hall arrays with superconducting interconnections,” *Metrologia*, vol. 56, no. 6, p. 065002, 2019.

[6] M. Kruskopf *et al.*, “Two-terminal and multi-terminal designs for next-generation quantized Hall resistance standards: contact material and geometry,” *IEEE Trans. Electron Dev.*, vol. 66, no. 9, pp. 3973 – 3977, 2019.

[7] A. F. Rigosi *et al.*, “Graphene devices for tabletop and high-current quantized Hall resistance standards,” *IEEE Trans. Instrum. Meas.*, vol. 68, pp. 1870 – 1878, 2019.

[8] F. P. M. Piquemal, *et al.*, “A first attempt to realize (multiple-QHE devices)-series array resistance standards,” *IEEE Trans. Instrum. Meas.*, vol. 48, no. 2, pp. 296 – 300, 1999.

[9] W. Poirier, *et al.*, “R_{K}/100 and R_{K}/200 Quantum Hall array resistance standards,” *J. Appl. Phys.* vol. 92, p. 2844, 2002.

[10] M. Ortolano, M. Abrate, and L. Callegaro, “On the synthesis of quantum Hall array resistance standards,” *Metrologia*, vol. 52, pp. 31 – 39, 2015.

[11] D.-H. Chae *et al.*, “Direct comparison of 1 MΩ quantized Hall array resistance and quantum Hall resistance standard,” *Metrologia*, vol. 55, pp. 645 – 652, 2018.

UNCERTAINTY IN THE SEISMIC RESPONSE OF REINFORCED CONCRETE STRUCTURES DUE TO MATERIAL VARIABILITY

C.L. Segura Jr.(1) and S. Sattar (2)

(1) Research Structural Engineer, National Institute of Standards and Technology, Gaithersburg, MD USA, christopher.segura@nist.gov
(2) Research Structural Engineer, National Institute of Standards and Technology, Gaithersburg, MD USA, siamak.sattar@nist.gov

Abstract

Inherent variability in the mechanical properties of reinforcing steel and concrete introduces uncertainty into the seismic assessment of reinforced concrete structures. Due to the high level of uncertainty associated with earthquake shaking characteristics, uncertainty due to variability in material properties is typically disregarded for seismic assessments. However, the potential impact of incorporating multiple sources of uncertainty in the seismic assessment framework is not well understood. To quantify the impact of material uncertainty, one-hundred iterations of a numerical model of a reinforced concrete structure are evaluated for their seismic performance at several earthquake shaking intensity levels. The one-hundred models differ only by the constitutive parameters used to model the materials, which are selected in accordance with measured statistical distributions of the mechanical properties of reinforcing steel bars and concrete. Material property statistical distributions, and correlations between material properties, are established using test data collected by the authors and test data available in the scientific literature. Preliminary results from analyses at the component level (i.e., individual column) indicate that dispersion in the predicted drift response for a given ground shaking intensity ($S_a[T_1]$) generally increases with shaking intensity, especially in the post-peak response regime for which a coefficient of variation (COV) in excess of 30 % is observed in the analyses presented herein. Of particular interest is the fact that a COV up to 20 % is observed for ground shaking intensities that produce only moderate ductility demands, prior to the onset of strength loss.

Keywords: uncertainty; performance-based earthquake engineering; seismic assessment; reinforced concrete

Segura, Christopher; Sattar, Siamak. "Uncertainty in the Seismic Response of Structural Components due to Material Variability." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Variability in the material properties of steel reinforcing bars and concrete affects the strength and deformation capacity of reinforced concrete structural components and, as a result, introduces uncertainty in the seismic assessment of reinforced concrete structures. Variability in material properties can be attributed to several factors including differences in the chemical composition of the steel and concrete materials, the source of the materials, and the material manufacturing processes. It is generally assumed that the contribution of material uncertainty to the overall uncertainty in the seismic assessment of a structure can be disregarded, particularly because of the relatively high level of uncertainty associated with earthquake shaking (record-to-record uncertainty). However, the potential impact of incorporating multiple sources of uncertainty, including that attributed to material variability, in the seismic assessment framework is not well understood. The importance of identifying and quantifying sources of uncertainty other than that associated with record-to-record variability has been highlighted by the results of recent blind prediction competitions in which relatively large dispersions have been reported for seismic response parameters predicted by experts in the field of nonlinear dynamic modeling of reinforced concrete structures [e.g., 1]. What is particularly concerning about the variability in contestant responses for blind prediction competitions is that the input earthquake shaking is known – that is, there is no record-to-record uncertainty.

The Performance-Based Earthquake Engineering (PBEE) framework [2,3] provides a convenient mechanism to account for the various sources of uncertainty that may impact the seismic assessment of structures. This paper describes the methodology used to quantify uncertainty due to variability in material properties. This work is one part of a project aimed at developing a framework to quantify the individual contributions of three main sources of uncertainty on the seismic response of reinforced concrete structures: 1) uncertainty associated with variability in material properties; 2) uncertainty related to the nonlinear modeling formulation used to conduct seismic analyses; and 3) uncertainty in earthquake shaking.

To evaluate the significance of material uncertainty, one-hundred iterations of a nonlinear structural analysis of a reinforced concrete structure are conducted at various earthquake shaking intensity levels. The one-hundred analysis iterations differ only by the constitutive parameters used to model the materials, which are selected in accordance with measured statistical distributions of the material properties of reinforcing steel and concrete, as well as correlations between material properties. To conduct the material uncertainty study, an analytical model representing a circular bridge column tested on the University of California, San Diego (UCSD) shake table in September 2010 [4] is developed. The UCSD bridge column structure is selected for uncertainty quantification because: 1) the bridge column is a simple structure that can be used to evaluate component-level uncertainty; 2) a blind prediction contest was organized and carried out by the UCSD research team [1] in coordination with the shake table test; and 3) comprehensive data are available on the experimental test.

The following sections describe: 1) the formulation of material statistical distributions used to conduct material uncertainty quantification; 2) the range of lateral drift predictions made by contestants of the UCSD blind prediction contest; 3) the development and verification of the analytical model of the bridge column; and 4) preliminary findings for the component-level (column) uncertainty study.

2. Statistical Variability in Material Properties

Variability in the mechanical properties of steel reinforcement and concrete is quantified as a set of statistical distributions and correlations developed using data available in the literature [5,6,7], an extensive database of steel reinforcing bar material tests provided to the authors courtesy of the Concrete Reinforcing Steel Institute (CRSI) [8], and a set of concrete cylinder tests collected by the authors. Material test data used by the authors consists of more than 80,000 tests on Grade 60 (f_y=420 MPa) reinforcing bars satisfying the requirements of ASTM A706 [9], as well as 88 concrete cylinder tests on normal-weight concrete with design compressive strengths between 31 MPa and 35 MPa. ASTM A706 places strict limits on the: 1) minimum and maximum allowable yield strength, to control the forces that can develop in yielding members; 2) minimum allowable tensile strength; 3) minimum allowable tensile-to-yield strength ratio, to ensure adequate spread of plasticity at yielding sections; and 4) minimum allowable tensile rupture strain.

Statistical distributions are developed for material parameters that describe the monotonic stress-strain behavior of Grade 60 ASTM A706 reinforcement and concrete materials, idealized in Fig. 1a and Fig. 1b, respectively. Important material properties include the yield strength (f_y), elastic modulus (E_s), strain hardening modulus (E_{sh}), tensile strength (f_u), and rupture strain (ε_{rup}) of reinforcing steel, as well as the peak compressive stress-strain (ε_{c0}, f_{c0}) and crushing strain (ε_{cu}) of concrete (Fig. 1).

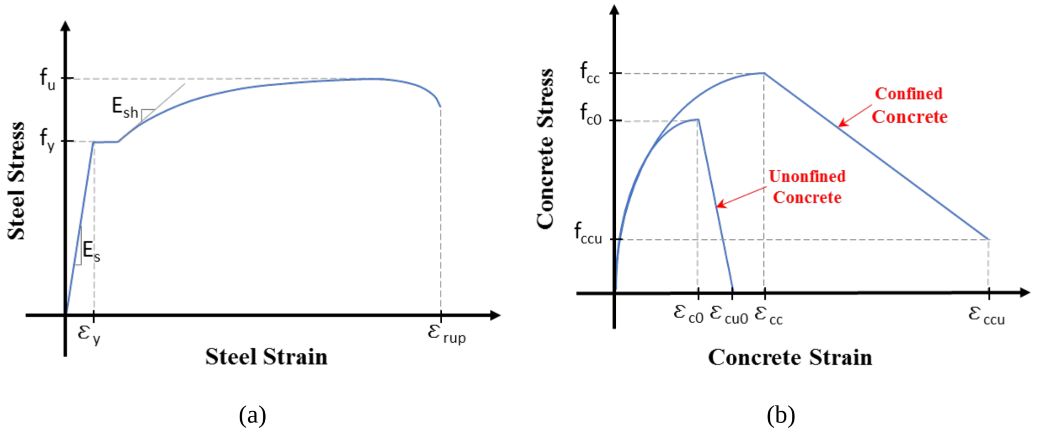
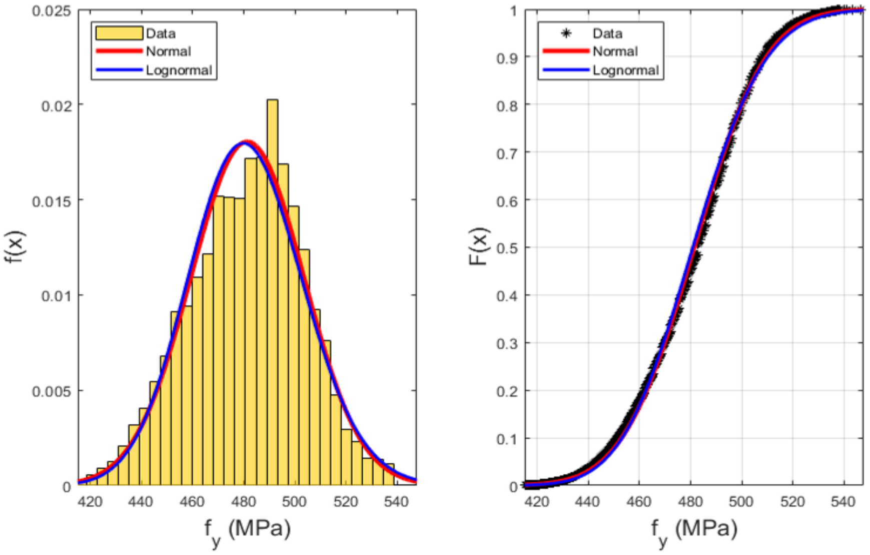


Fig. 1 – Theoretical Stress vs. Strain Behavior – (a) Reinforcing Steel in Tension; and (b) Concrete in Compression (Compressive Stress Shown as Positive)

For uncertainty quantification, material properties are represented as a set of statistical distributions. Fig. 2 provides a representative comparison of the experimental data to normal and lognormal probability distribution function (PDFs) (Fig. 2a) and cumulative distribution functions (CDFs) (Fig. 2b). Correlation coefficients between material properties are determined using a Spearman correlation analysis with a 95 % confidence criterion [10].

Confined concrete properties are not easily derived from test data because several factors contribute to the confined properties. However, confinement models that have been developed using laboratory tests that account for the various factors that affect the confined properties are available in the literature. Uncertainty in confined properties of concrete is quantified using the confinement models developed by Saatcioglu and Razvi [11] and Legeron and Paultre [12]. Specifically, statistical distributions for the peak confined stress-strain (ε_{cc}, f_{cc}) and confined concrete crushing strain (ε_{ccu}) are determined. Two additional confinement models [13,14] were evaluated, but both were deemed to introduce excessive dispersion in confined properties because the model was either developed specifically for rectangular sections, while the analyses reported herein are for a circular column, or because the method used to determine strains on the softening branch of the confined stress-strain curve is inconsistent with the other confinement models [15].



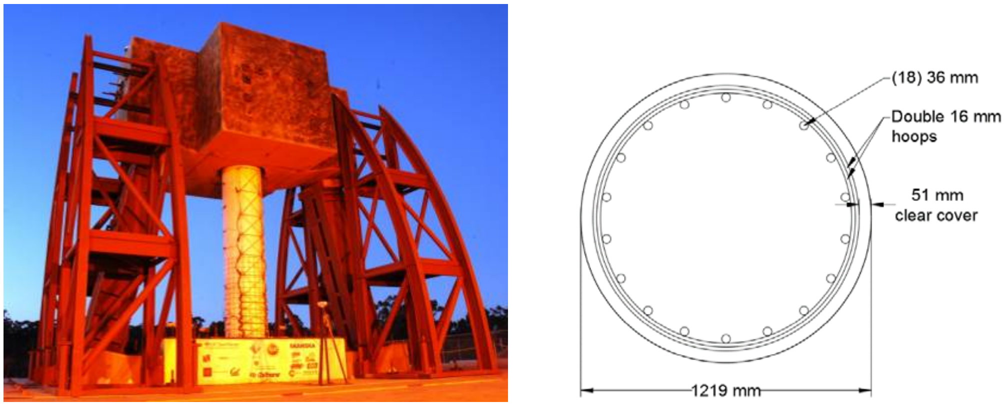
(a) (b)

Fig. 2 – Measured Reinforcing Steel Yield Strength (f_y) – (a) Comparison of Histogram to PDF; and (b) Comparison of Measured Values to CDF

3. Model Development and Verification

3.1 UCSD Shake Table Test and Blind Prediction Contest

The impact of material uncertainty is quantified by evaluating the seismic performance of an analytical model representing a circular bridge column tested on the University of California, San Diego (UCSD) shake table in September 2010 [4]. A photo of the shake table test setup and a cross-sectional drawing of the column are provided in Fig. 3. The column was 1219 mm in diameter and 7315 mm in height. A large concrete mass with

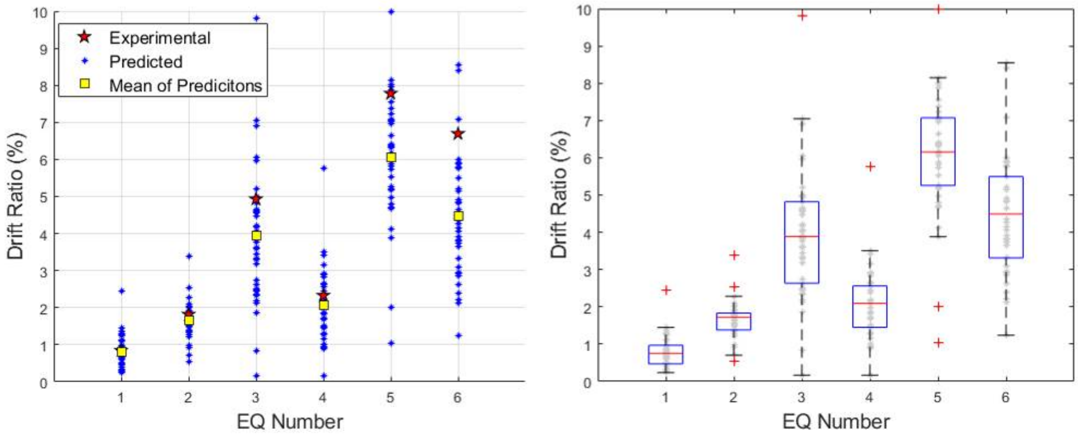


(a) (b)

Fig. 3 – UCSD Bridge Column Shake Table Test – (a) Test Setup; and (b) Column Cross-Section [4]

a total tributary weight of 2.32 MN was attached to the top of the column (Fig. 3). Longitudinal reinforcement consisted of eighteen 36 mm diameter bars spaced evenly about the circumference of the column. Transverse reinforcement consisted of two bundled 16 mm diameter hoops spaced at 152 mm on-center. All reinforcement was specified as Grade 60 (f_y=414 MPa) ASTM A706 and the specified concrete compressive strength (f'_c) was 28 MPa. Measured concrete cylinder compressive strength at the time of testing was 41 MPa. Compressive strains at peak cylinder strength ranged between about 0.0025 and 0.003. The average yield strength and ultimate strength measured for two of the 36 mm diameter reinforcing bar samples were 518 MPa and 706 MPa, respectively, and average rupture strain for the rebar samples was 0.122 [4].

The bridge column was subjected to a series of six earthquake acceleration records, designated as EQ1 through EQ6, that were selected and scaled for target displacement ductility demands of 1.0 (EQ1), 2.0 (EQ2 and EQ4), 4.0 (EQ3 and EQ6), and 8.0 (EQ5). EQ2 and EQ4 used the same scaled record, as did EQ3 and EQ6 [1,4]. Fig. 4a compares maximum drift ratios predicted by the forty-two participants of the blind prediction contest to the experimentally measured responses for the six applied ground motion records. The mean of the predicted responses is also plotted in Fig. 4a. The relative difference in the mean predicted response and the experimentally measured drift ratio (i.e., mean bias) ranges between -6 % (EQ1) and -33 % (EQ6), indicating that the analytical models, on average, tend to underestimate the deformation of the bridge column. Fig. 4b presents boxplots of the contestant predictions for each earthquake record. For each boxplot, the bottom and top edges of the "box" indicate the extents of the data within one standard deviation of the mean of the predicted responses; the top and bottom "whiskers" indicate data within 2.7 standard deviations of the mean; and the "+" markers indicate outliers from a normal distribution. Relatively large dispersion in the predicted response is evident – particularly for EQ3, EQ5 and EQ6 – demonstrating the importance of quantifying potential sources of uncertainty in the seismic assessment framework.



(a) (b)

Fig. 4 – UCSD Blind Prediction Contest Responses – (a) Comparison of Predicted Drift Ratio to Experimental and Mean; and (b) Boxplot of Predicted Responses

3.2 Analytical Model Description

A distributed plasticity model (i.e., fiber model) of the bridge column is developed and nonlinear analyses are conducted in OpenSees [16]. A fiber element formulation is selected because, unlike lumped plasticity models, fiber models enable direct definition of material constitutive properties, thereby making it possible to straightforwardly quantify uncertainty due to variability in material properties. A displacement-based fiber element model (stiffness formulation) is developed and verified against experimental measurements from the UCSD shake table test. The spatial discretization of the fiber model is shown in Fig. 5. The model consists of eight equal length elements, each with three Gauss-Lobatto integration points (Fig. 5a) [17]. An axial load of 2.32 MN is applied at the top node of the column model (Fig. 5a) and held constant throughout the analyses. A lumped mass of 237 000 kg (i.e., 2.32 MN/g_a) is also applied at the top node of the column. Confined concrete is discretized into 8 sections in the radial dimension of the column and 8 sections in the circular direction for a total of 64 confined concrete fibers (Fig. 5b). A total of 32 fibers (4 radial, 8 circular) are used to model unconfined cover concrete. Steel reinforcing bars are discretized into 18 fibers located at the centroid of the bar locations indicated in Fig. 3. Second-order P-Delta effects are accounted for in the nonlinear analyses.

The Concrete02 constitutive model implemented in OpenSees is used to model unconfined concrete. Confined concrete is modeled using the Concrete07 constitutive model and steel reinforcing bar materials are modeled using the SteelMPF model, both of which employ sophisticated constitutive hysteretic rules [18,19,20]. Regularization of the compressive material properties for concrete and steel is conducted in accordance with the technique developed by Coleman and Spacone [21]. The regularization technique adjusts the stress-strain relationships for uniaxial fibers such that analytical results are insensitive to the model spatial discretization. The OpenSees MinMax constitutive model is used to assign tension and compression strain limits for the SteelMPF reinforcing bar constitutive model. The tension strain limit is set as the steel rupture strain capacity while the compression strain limit signals the reinforcing bar material to degrade to zero stress (i.e., rebar buckling) once the confined concrete reaches its crushing strain.

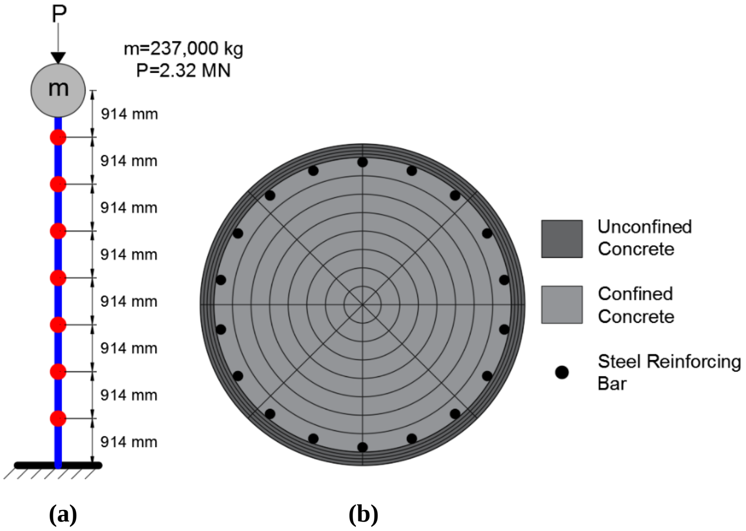


Fig. 5 – Analytical Model Discretization – (a) Elements and Nodes; and (b) Cross-Section

3.3 Model Verification

Model verification is conducted by comparing the response of the analytical model to the response measured experimentally on the shake table at UCSD. To do so, the analytical model is subjected to the acceleration time series applied during the UCSD shake table test, which consisted of six earthquake acceleration records applied separately. A comparison of the analytical and experimentally measured lateral drift ratio, measured at the top of the column, is shown in Fig. 6. Only the experimental deformation attributed to flexure is plotted in Fig. 6 to enable a direct comparison with the results for the fiber element model, which is only capable of capturing the flexural response of the column. It is evident in Fig. 6 that the analytical model is capable of capturing the deformation response of the bridge column, including residual inelastic deformations. For each earthquake record, the maximum analytical and experimental drift ratio are compared in Fig. 6. The difference in the predicted and experimental maximum drift ratio ranges between 3 % (EQ3) and 29 % (EQ5). It is emphasized that model parameters are not "calibrated" to achieve the results presented in Fig. 6; rather, the uniaxial fiber material parameters are taken directly from the measured material properties [4] and confinement models, adjusted for mesh size in accordance with the material regularization technique developed by Coleman and Spacone [21].

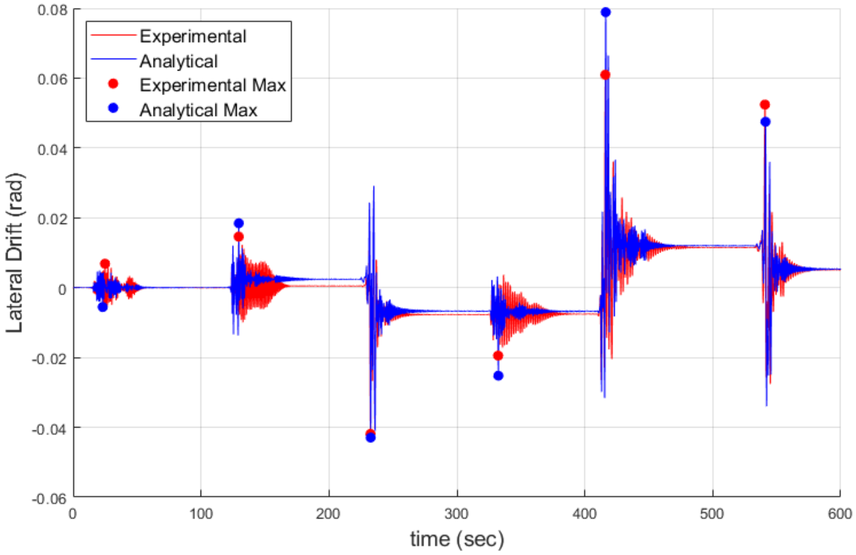


Fig. 6 – Model Verification – Experimental vs. Analytical Lateral Drift Ratio

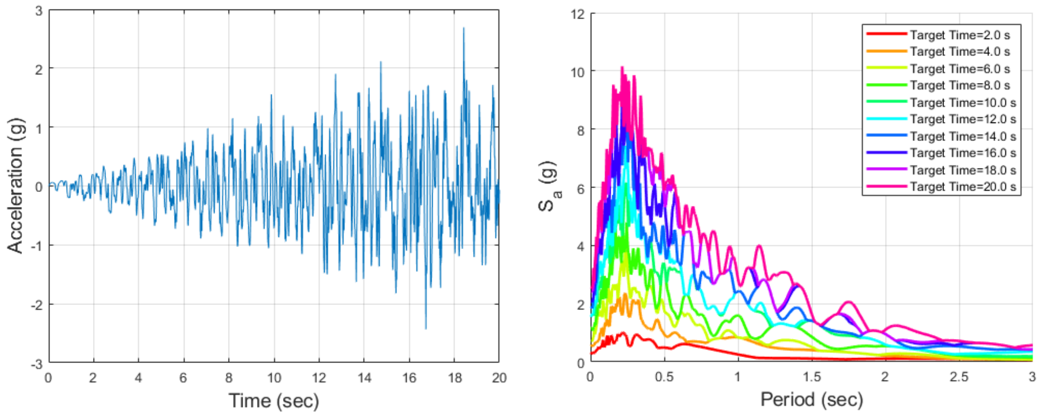
4. Uncertainty Quantification

One-hundred iterations of the analytical model (Fig. 5) are constructed and nonlinear analyses are conducted in OpenSees to quantify the impact of uncertainty in the material properties of reinforcing steel and concrete. The one-hundred iterations of the analytical model are the same in their formulation (Fig. 5), differing only by the material properties used to define the constitutive behavior of the fiber elements. The input constitutive parameters for each model are selected from a set of one-hundred realizations of reinforcing steel, unconfined concrete, and confined concrete material properties. Mean, minimum, and maximum values of the one-hundred material realization set are given in Table 1 for three representative material properties (f_{c0}, f_y, and ε_{rup}). The one-hundred material property realizations are selected to represent the statistical distributions of the material properties, and correlations between properties, described in Section 2 with minimal variation from the desired statistical distributions. The variation in the mean and coefficient of variation for the one-hundred material realization set and the statistical distributions is below 5 % for all material properties.

Table 1: Ranges of Representative Material Properties (See Fig. 1 for Description of Properties)

	Mean	Minimum	Maximum
$\mathbf{f_{c0}}$	33.8 MPa	23.5 MPa	46.4 MPa
$\mathbf{f_y}$	482 MPa	429MPa	529 MPa
$\boldsymbol{\varepsilon_{rup}}$	0.157	0.107	0.211

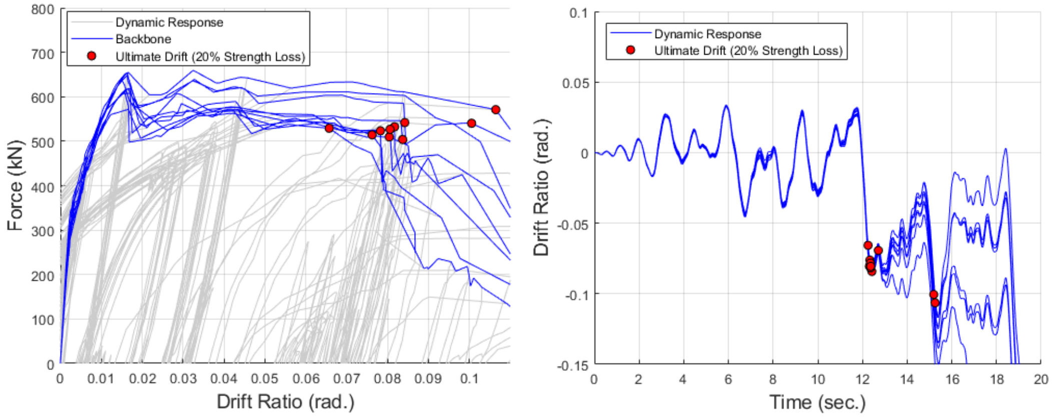
Nonlinear analyses of the analytical models are conducted in OpenSees using an Endurance Time Acceleration Function (ETAF) (e.g., see [22]). ETAF enables the evaluation of the analytical model at progressively increasing ground shaking intensities using a single, unique ground acceleration record. ETAF is, therefore, analogous to a dynamic pushover analysis. A plot of ground acceleration vs. time for the ETAF is provided in Fig. 7a. Fig. 7b presents the 2.5 %-damped pseudo acceleration response spectra (S_a) for ten different target time intervals. A target time interval represents the elapsed time from the beginning of the record. For example, a target time of 10 seconds captures ground shaking for the time between 0 and 10 seconds, whereas a target time of 12 seconds captures ground shaking from time 0 to 12 seconds. As shown in Fig. 7b, the spectral acceleration demands for the ETAF record increase with increasing target time.



(a) (b)

Fig. 7 – (a) Acceleration Record for ETAF; and (b) 2.5 %-Damped Response Spectra (S_a) at Various Target Times

Fig. 8a presents the lateral force vs. drift ratio backbone curves for ten representative analyses of the one-hundred model iterations. The models demonstrate very ductile response, as would be expected for the well-confined section with relatively low axial stress. For each model, the ultimate drift ratio from the analysis is indicated as a red circle on the figure. The ultimate drift ratio is designated as the lateral drift at the last analysis step for which the residual strength exceeds 80 % of the peak capacity (i.e., 20 % strength loss). Variability in the column shear force is evident beyond a drift ratio of about 1 %; however, as shown in Fig. 8b, dispersion in the deformation response of the column is small until the onset of strength loss. Following strength loss, a wide range of deformation responses is evident. The coefficient of variation of the predicted lateral drift ratio ranges between 2.5 % and 8 % for lower earthquake shaking intensities – that is, for target times less than about 8 seconds. For higher earthquake shaking intensities, following strength loss, coefficients of variations reach approximately 16 %.



(a) (b)

Fig. 8 – (a) Lateral Force vs. Drift Ratio; and (b) Drift Ratio Time Series

Overall, dispersion in the predicted deformation response of the column for a particular analysis target time is within what is typically considered an acceptable range for seismic analysis [e.g., see 23,24]. It is important to point out, however, that in the PBEE framework [2,3], the vulnerability of a structure to collapse is often expressed in terms of the probability of reaching a particular engineering demand parameter (EDP) at a given ground shaking intensity. The EDP and the ground shaking intensity measure often used to characterize collapse risk are the maximum lateral drift (Δ_{max}) and the first mode elastic spectral acceleration ($S_a[T_1]$), respectively. Characterization of the dispersion in Δ_{max} for a given $S_a(T_1)$ is, therefore, important to account for uncertainty in the PBEE framework. To illustrate this dispersion, the predicted Δ_{max} values determined for the one-hundred model iterations are presented as lognormal PDFs in Fig. 9 for three earthquake shaking intensity levels: $S_a(T_1)$=0.5g, $S_a(T_1)$=1.5g, and $S_a(T_1)$=3.0g. The best-fit distribution based on the Kolmogorov–Smirnov test at different shaking intensities may differ. A lognormal distribution is used herein because it is commonly used in the field of structural engineering to represent the uncertainty in the lateral drift response of structures. A modal damping coefficient of 2.5 % is assumed for the spectral accelerations in Fig. 9, as is the case for the spectra shown in Fig. 7b. As can be seen in Fig. 9, dispersion generally increases for increasing $S_a(T_1)$. For smaller spectral acceleration demands (i.e., $S_a[T_1]$=0.5g), the dispersion in drift ratio is low. This is expected as the column response is primarily elastic for this level of shaking. As yielding of the column's longitudinal reinforcement occurs (i.e., $S_a[T_1]$=1.5g), dispersion in Δ_{max} increases. Strength loss is observed in most of the one-hundred models for $S_a[T_1]$>2g, after which dispersion in the predicted drift response becomes large in proportion to the variability in $S_a(T_1)$. A similar trend of increasing dispersion in the lateral drift response for higher earthquake shaking intensity levels has been observed for steel columns [25]. It is noted that, due to the use of the ETAF loading protocol (Fig. 7), drift ratio results for larger $S_a(T_1)$ values include previous damage associated with lower $S_a(T_1)$ values.

Table 2 summarizes statistical parameters for the lognormally distributed Δ_{max} at the three ground shaking intensities shown in Fig. 9. The mean $(\bar{\mu}_{\Delta_{max}})$ and median $(\tilde{\mu}_{\Delta_{max}})$ values reported in Table 2 are calculated according to Eqns. 1 and 2, respectively:

$$\bar{\mu}_{\Delta_{max}} = e^{\mu_{\ln(x)}+\frac{1}{2}\sigma_{\ln(x)}{}^2}, \tag{1}$$

$$\tilde{\mu}_{\Delta_{max}} = e^{\mu_{\ln(x)}}, \tag{2}$$

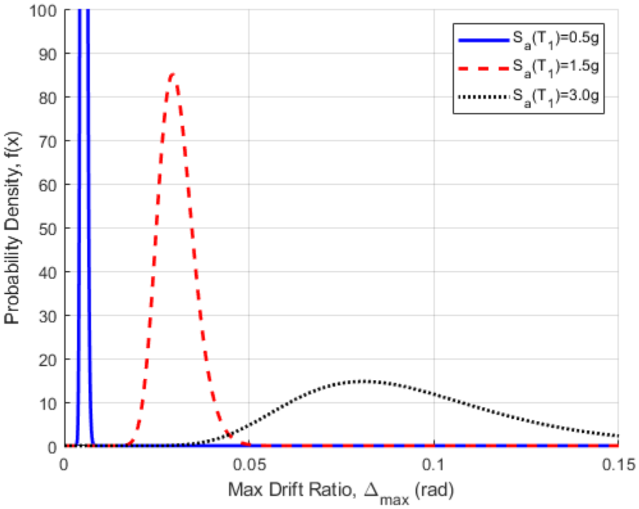


Fig. 9 – PDF of Maximum Drift Ratio for Different $S_a(T_1)$

where $\mu_{\ln(x)}$ and $\sigma_{\ln(x)}$ are the mean and standard deviation of the natural logarithms of drift values. The coefficient of variation reported in Table 2 is calculated according to Eqn. 3:

$$COV_{\Delta_{\max}} = \sqrt{e^{\sigma_{\ln(x)}^{2}} - 1}. \tag{3}$$

For ground shaking intensities that result in elastic response of the column model (e.g., $S_a[T_1]$=0.5g), the coefficient of variation for Δ_{max} is generally around 10 % or less. In the inelastic response regime, prior to the onset of strength loss (e.g., $S_a[T_1]$=1.5g), the dispersion in Δ_{max} becomes larger, reaching a coefficient of variation up to approximately 20 %. The largest dispersion in the predicted drift response (e.g., COV≈30 % for $S_a[T_1]$=3.0g) is associated with a reduction in the lateral load-carrying capacity of the column (softening). The larger dispersion in the softening regime is attributed to the uncertainty in the post-peak behavior of the confined concrete material.

Table 2: Statistical Parameters of Δ_{max} for Different $S_a(T_1)$

$S_a(T_1)$	Mean $(\bar{\mu}_{\Delta_{\max}})$	Median $(\tilde{\mu}_{\Delta_{\max}})$	Coefficient of Variation $(COV_{\Delta_{\max}})$
0.5g	0.53 %	0.52 %	10.8 %
1.5g	3.04 %	3.01 %	15.9 %
3.0g	9.42 %	8.95 %	32.7 %

5. Summary

A methodology to quantify uncertainty in the seismic assessment of structures due to variability in the mechanical properties of concrete and reinforcing steel is presented. This material uncertainty quantification study is one part of an ongoing project aimed at developing a framework to quantify the individual contributions of uncertainty on the seismic response of reinforced concrete structures, at both the component level (individual column) and at the system level (frame) due to: (1) variability in material properties, (2) the choice of nonlinear modeling formulation used to conduct seismic analyses, (3) and earthquake shaking. Uncertainty in the seismic response of a structural component is quantified using one-hundred iterations of an analytical model of a reinforced bridge concrete column. The one-hundred model iterations capture the statistical distributions of concrete and reinforcing steel mechanical properties determined using data collected by the authors and data available in the scientific literature. Analytical seismic responses at various earthquake shaking intensities (i.e., $S_a[T_1]$) are determined for the one-hundred models using an Endurance Time Acceleration Function, which subjects the model to increasing seismic demands using a single ground acceleration record. Results of the component level study demonstrate that dispersion in the predicted drift response due to material variability generally increases with increasing $S_a(T_1)$. In the elastic response regime, dispersion is low (coefficient of variation, COV, around 10 % or less); however, as flexural yielding occurs, dispersion increases (COV up to 20 %) even though a hardening response is still evident in the models. Larger dispersion in the post-peak regime (COV around 30 %) can be attributed to greater uncertainties in the material response of concrete as material softening occurs.

6. Acknowledgements

The authors would like to express their gratitude to the Concrete Reinforcing Steel Institute (CRSI) for providing reinforcing bar mechanical property test results collected from steel reinforcing bar manufacturers in North America. The authors would also like to thank Dr. Vesna Terzic for providing the response data for the UCSD blind prediction contest.

7. References

[1] Terzic, V., Schoettler, M.J., Restrepo, J.I., and Mahin, S.A. (2015): Concrete Column Blind Prediction Contest 2010: Outcomes and Observations. *Technical Report PEER 2015/01*, Pacific Earthquake Engineering Research, Berkeley, USA.

[2] Porter, K.A. (2003): An Overview of PEER's Performance-Based Earthquake Engineering Methodology. *Ninth International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP9)*, San Francisco, USA.

[3] Moehle, J. and Deierlein, G.G. (2004): A Framework Methodology for Performance-Based Earthquake Engineering. *13th World Conference on Earthquake Engineering*, Vancouver, Canada.

[4] Schoettler, M.J., Restrepo, J.I., Guerrini, G., Duck, D.E., and Carrea, F. (2015): A Full-Scale, Single-Column Bridge Bent Tested by Shake Table Excitation. *Technical Report PEER 2015/02*, Pacific Earthquake Engineering Research, Berkeley, USA.

[5] Mirza, S.A. and MacGregor, J.G. (1979): Variability of Mechanical Properties of Reinforcing Bars. *ASCE Journal of the Structural Division*, **105** (5), 921-937.

[6] Mirza, S.A., Hatzinikolas, M., and MacGregor, J.G. (1979): Statistical Description of Strength of Concrete. *ASCE Journal of the Structural Division*, **105** (6), 1021-1037.

[7] Nowak, A.S., Szersen, M.M., Szeliga, E.K., Szwed, A., and Podhorecki, P.J. (2008): Reliability-Based Calibration for Structural Concrete, Phase 3. *Report to the Portland Cement Association and Precast/Prestressed Concrete Institute*, Skokie, USA.

[8] CRSI (2018): CRSI Mill Database. *Annual Summary Reports for 2011-2017*. Concrete Reinforcing Steel Institute (CRSI), Schaumburg, USA.

[9] ASTM A706/A706M (2016): Standard Specification for Deformed and Plain Low-Alloy Steel Bars for Concrete Reinforcement. American Society for Testing and Materials (ASTM) International, West Conshohocken, USA.

[10] Spearman, C. (1904): The Proof and Measurement of Association Between Two Things. *The American Journal of Psychology*, **15** (1), 72-101.

[11] Saatcioglu, M. and Razvi, S.R. (1992): Strength and Ductility of Confined Concrete. *ASCE Journal of Structural Engineering*, **118** (6), 1590-1607.

[12] Legeron, F. and Paultre, R. (2003): Uniaxial Confinement Model for Normal- and High-Strength Concrete Columns. *ASCE Journal of Structural Engineering*, **129** (2), 241-252.

[13] Mander, J.B., Priestley, M.J.N., and Park, R. (1988): Theoretical Stress-Strain Model for Confined Concrete. *ASCE Journal of Structural Engineering*, **114** (8), 1804-1826.

[14] Scott, B.D., Park, R., and Priestley, M.J.N. (1982): Stress-Strain Behavior of Concrete Confined by Overlapping Hoops at Low and High Strain Rates. *ACI Journal*, **79** (1), 13-27.

[15] Arteta, C. A. (2015): Seismic Response Assessment of Thin Boundary Elements of Special Concrete Shear Walls. *PhD Dissertation*, University of California, Berkeley, USA.

[16] Mazzoni, S., McKenna, F., and Fenves, G.L. (2006): OpenSees Command Language Manual, University of California, Berkeley, Berkeley, USA.

[17] Abramowitz, M. and Stegun, I.A. (1964): Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables. *National Bureau of Standards Applied Mathematics Series*, National Bureau of Standards, Washington D.C., USA.

[18] Chang, G.A. and Mander, J.B. (1994): Seismic Energy Based Fatigue Damage Analysis of Bridge Columns: Part 1 – Evaluation of Seismic Capacity. *Technical Report for the National Center for Earthquake Engineering Research*. The University at Buffalo, State University of New York, Buffalo, USA.

[19] Menegotto, M., and Pinto, P.E. (1973): Method of Analysis of Cyclically Loaded RC Plane Frames Including Changes in Geometry and Non-Elastic Behavior of Elements Under Normal Force and Bending. *Symposium on Resistance and Ultimate Deformability of Structures Acted on by Well-Defined Repeated Loads*. Lisbon, Portugal.

[20] Filippou F.C., Popov, E.P., and Bertero, V.V. (1983): Effects of Bond Deterioration on Hysteretic Behavior of Reinforced Concrete Joints. *Report No. UCB/EERC-83/19*. Earthquake Engineering Research Center, University of California, Berkeley, Berkeley, USA.

[21] Coleman, J. and Spacone, E. (2001): Localization Issues in Force-Based Frame Elements. *ASCE Journal of Structural Engineering*, **127** (11), 1257-1265.

[22] Hariri-Ardebili, M.A., Sattar, S., and H. E. Estekanchi (2014): Performance-based Seismic Assessment of Steel Frames Using Endurance Time Analysis. *Engineering Structures*, **69**, 216-234.

[23] FEMA (2009): Quantification of Building Seismic Performance Factors (FEMA P695). *FEMA P695*, Prepared by the Applied Technology Council for the Federal Emergency Management Agency (FEMA), FEMA, Washington D.C., USA.

[24] FEMA (2000): Recommended Seismic Design Criteria for New Steel Moment-Frame Buildings (FEMA 350). *FEMA 350*, Prepared by the SAC Joint Venture for the Federal Emergency Management Agency (FEMA), FEMA, Washington D.C., USA.

[25] Sattar, S., Weigand, J.M, and Wong, K.K.F. (2018): Quantification of Uncertainties in the Response of Beam-Columns in Steel Moment Frames. *11th U.S. National Conference on Earthquake Engineering*, Los Angeles, USA.

Josephson Arbitrary Waveform Synthesizer as a Reference Standard for the Calibration of Lock-in Amplifiers

D Georgakopoulos*, I Budovsky*, and S. P. Benz†

*National Measurement Institute Australia, 36 Bradfield Road, Lindfield NSW 2070, Australia
dimitrios.georgakopoulos@measurement.gov.au

†National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305, USA

***Abstract* — We have extended the voltage range of the Josephson arbitrary waveform synthesizer from 1 mV down to 1 µV at frequencies from 60 Hz to 1 kHz to calibrate precision lock-in amplifiers. Experimental results show that the system's uncertainty is dominated by the resolution of the lock-in amplifier. We anticipate that the Josephson arbitrary waveform synthesizer-based system will extend the lower voltage and frequency range of ac voltage metrology and improve the uncertainties by one order of magnitude compared to conventional techniques.**

***Index Terms* — Voltage measurement, voltage standards, electric potential, Josephson junction, measurement uncertainty.**

I. Introduction

A Josephson arbitrary waveform synthesizer (JAWS) utilizes the accuracy of the ac Josephson effect to produce voltages calculated from first principles at frequencies up to several megahertz [1]. Behr *et al.* [2] used the JAWS to generate small voltages (from some nV to some mV). We have extended the use of a JAWS to the voltage range from 1 µV to 1 mV, at frequencies from 60 Hz to 1 kHz, to provide traceability to lock-in amplifiers. This offers a convenient and intrinsically accurate alternative to conventional methods for the calibration of lock-in amplifiers, such as those based on the generation of low voltages by means of a calibrated semiconductor source and a calibrated inductive voltage divider (IVD) [3].

II. System Description

Fig. 1 shows the block diagram of the JAWS system. The system is described in [4] and uses Josephson junction arrays (JJAs) and a microwave circuit design developed at the National Institute of Standards and Technology (NIST) [5]. The low voltages required to calibrate lock-in amplifiers are generated directly by the JJAs without any compensation signal. An arbitrary waveform generator (AWG) is used to provide an external reference voltage to phase-lock the lock-in amplifier.

III. Experimental Results

The JAWS produces quantized voltage pulses with a time integral of *h/2e* across each Josephson junction. Given that the JJA produces quantized pulses, for the JAWS to be a fundamental standard its operation must be independent of (a) the JJA used; (b) the biasing electronics (rf amplifiers and dc blocks); (c) the rf power; and, (d) the repetition frequency and the rf pattern of the pulses (i.e., if a particular waveform can be obtained by using more than one repetition frequency and pattern of pulses, the resulting waveforms should be the same). The various parts of the JJA generate sinewaves with slightly different phases due to delays in the arrival of the pulses within various parts of the JJA or the Josephson junctions, resulting in a voltage error. The last condition (d) tests for any errors due to these delays since the number of changes from a "zero" to a "non-zero" value (and vice versa) in the pattern depends on the repetition frequency.

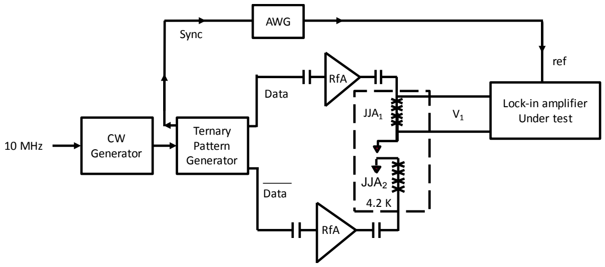


Fig. 1. Block diagram of the JAWS-based system for the calibration of lock-in amplifiers.

We conducted several experiments to provide evidence that these conditions are met. In this summary, we present some of these experiments. We used a commercial lock-in amplifier (SR865A) as a transfer standard.[1] Further experiments will be presented at the conference.

First, we investigated whether the operation of the JAWS-based system was independent of the JJA. We used the same biasing electronics (ternary pattern generator and rf amplifier) to sequentially drive two different JJAs and measured the correction of the lock-in amplifier (1 mV at 1 kHz). The results agreed within 0.08 µV, well below the resolution of the lock-in amplifier (0.1 µV for the 1 mV range).

Next, we investigated whether the JAWS system is independent of the biasing electronics and the repetition

[1] Commercial instruments and design tools are identified in this paper to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST and NMIA, nor does it imply that the equipment identified is necessarily the best available for the purpose.

Georgakopoulos, Dimitrios; Budovsky, Ilya; Benz, Samuel P. "Josephson Arbitrary Waveform Synthesizer as a Reference Standard for the Calibration of Lock-in Amplifiers." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

frequency of the rf pattern of pulses used to generate the desired signal. We generated four waveforms with the same rms value (1 mV) and frequency (1 kHz), using four different rf repetition frequencies. We then used two different sets of biasing electronics (CH1 and CH2) to bias the same JJA sequentially and measured the correction of the lock-in amplifier (Fig. 2). The results show consistency in the measurements.

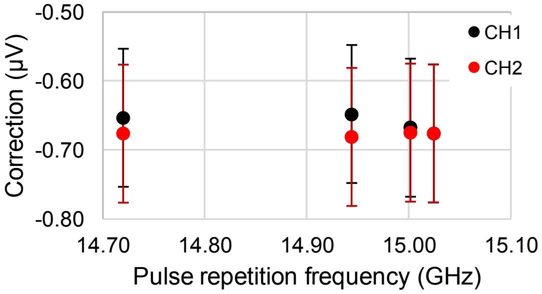


Fig. 2. Correction of the lock-in amplifier at 1 mV and 1 kHz for various pulse repetition frequencies. The error bars show the resolution of the lock-in amplifier

The quantization of the JAWS is usually checked by measuring the spectrum of the generated signal. Fig. 3 shows the spectra of these waveforms, measured with a National Instruments 5922 digitizer at a sampling frequency of 10 MHz, and the spectrum of a signal with the same characteristics produced by dividing a 1 V output voltage of a calibrator (Fluke 5700A) with an IVD (Sullivan F9200) with ratio 1000:1. Fig. 3 also shows the Fourier transform of one of the pattern of pulses downloaded to the pattern generator (for repetition frequency 15.0016 GHz) and the ideal spectrum of a Δ-Σ modulator with similar characteristics to the one used for the generation of the pattern of pulses. As can be seen from Fig. 3, the noise floor and the spurious tones of the JAWS and the calibrator/IVD signals are similar, which suggests that the various spurious tones and the noise floor in the spectrum are features of the digitizer.

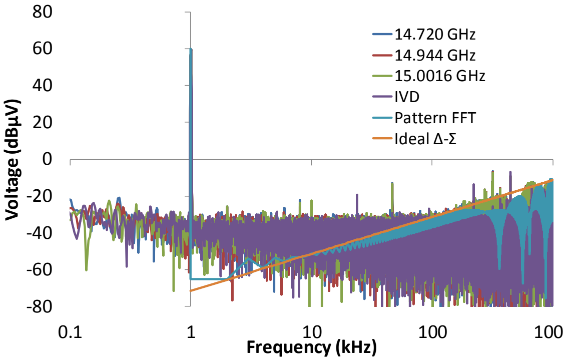


Fig. 3. Spectra of the four waveforms (1 mV at 1 kHz) produced by different means (see text for details).

The corrections of the lock-in amplifier were measured at a number of voltages in the range from 1 µV to 1 mV and at frequencies of 60 Hz, 200 Hz and 1 kHz. Figs. 4a and 4b show the results for 1 µV and 1 mV, respectively. For comparison, Figs. 4a and 4b also show the corrections of the same lock-in amplifier measured with a conventional lock-in amplifier calibration system consisting of a semiconductor source (Fluke 5700A) and an IVD (Sullivan 9700A). The results agree within the resolution of the lock-in amplifier.

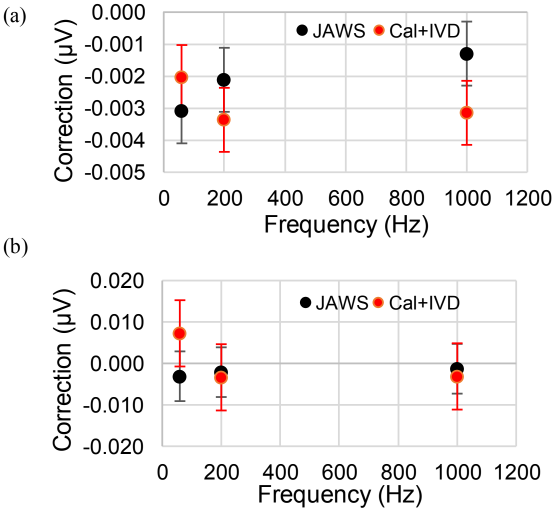


Fig. 4. Correction of the lock-in amplifier at (a) 1 µV and (b) 1 mV for different frequencies. The error bars show the resolution of the lock-in amplifier (1 nV and 0.1 µV, respectively) for the corresponding ranges.

Conclusion

The results presented in this summary show that the JAWS gives consistent results for the generation of low voltages from 1 µV to 1 mV at frequencies up to 1 kHz. The uncertainty analysis, to be presented at the conference, suggests that the dominant uncertainty component is the resolution of the lock-in amplifier under test.

References

[1] J.A. Brevik *et al.*, "Josephson arbitrary waveform synthesis with multi-level pulse biasing," *IEEE Trans. Appl. Supercond.,* vol. 27, no. 3, Apr. 2017, Art. No. 1301707.

[2] Behr *et al.* "Precision µV-synthesizer based on a pulse-driven Josephson array," CPEM 2016 Conf. Digest, p. 1, June 2016.

[3] D. Corminboeuf, "Calibration of the absolute linearity of lock-in amplifiers," *IEEE Trans. Instrum. Meas.,* vol. 68, no. 6, pp. 2060 – 2065, June 2019.

[4] D. Georgakopoulos *et al.*, "Josephson arbitrary waveform synthesizer as a reference standard for the measurement of the phase of harmonics in distorted waveforms," *IEEE Trans. Instrum. Meas.,* vol. 68, no. 6, pp. 1927 – 1934, June 2019.

[5] S.P. Benz *et al.*, "One-volt Josephson arbitrary waveform synthesizer," *IEEE Trans. Appl. Supercond.,* vol. 25, no. 1, Feb. 2015, Art. No. 1300108.

Polarization-dependent absolute-phase-corrected multidimensional coherent spectra of exciton-polaritons

Jagannath Paul[1,2], Jared K. Wahlstrand[1], and Alan D. Bristow[1,2]

[1]Nanoscale Device Characterization Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]Department of Physics & Astronomy, West Virginia University, Morgantown, WV 26506-6315, USA

ABSTRACT

Multidimensional coherent spectroscopy measures the third-order polarization response of a system to reveal microscopic electronic and many-body phenomena. Applied to semiconductor nanostructures, it can distinguish homogeneous and inhomogeneous broadening due to disorder or strain gradients, resolve coupling between transitions, and optically access transitions that are either non-radiating or outside the bandwidth of the pulses. Two tools often exploited in this versatile technique are (i) the ability to control the polarization of the excitation and emission and thus the optical selection rules, and (ii) the ability to capture the complex spectrum. Here, the polarization of pulses emerging from a multidimensional optical nonlinear spectrometer (MONSTR) and the resulting four-wave mixing emission are controlled automatically using variable retarders, such that multiple spectra are recorded during a single phase-stabilized scan. This improves the acquisition time by $\sim 3\times$ compared to running separate polarization scans. Importantly, only one phase ambiguity exists in the complex spectra across all sets of polarization states measured. This single ambiguity is resolved by comparing the initial spectrally resolved transient absorption to the complex four-wave mixing spectrum for collinear polarization and then applying it to all spectra. Here, the method is applied to a quantum well embedded in a semiconductor microcavity with an adjustable cavity-exciton detuning. The complex 2DCS spectra we report constitute the first measurements of detuning- and polarization-dependent exciton-polariton lineshape across the strong coupling regime.

Keywords: Multidimensional coherent spectroscopy, Coherent coupling of Exciton-Polariton, Phase corrected 2D Spectra, Four wave mixing

1. INTRODUCTION

Optical properties of semiconductor quantum wells (QW) are dominated by excitonic resonances at low temperature. Excitons are Coulomb bound electron-hole pairs. When such a QW is placed inside a planar semiconductor microcavity at the antinode of the resonant field of the cavity, the QW exciton strongly interacts with the cavity normal mode, resulting in the formation of a new quasiparticle, exciton-polaritons.[1] The single exciton line is replaced with lower polariton (LP) and upper polariton (UP) branches, with a splitting characterized by the vacuum Rabi splitting energy E_{VRS}. Exciton-polaritons are hybrid modes composed of an exciton-like part and a light-like cavity photon mode part. The energies of the LP and UP branches and how these parts are divided between them depends on the detuning (Δ) of the cavity mode with respect to the exciton energy. The detuning-dependent branch energies are depicted in Fig. 1(a). The exciton and photon fractions of the UP or LP branches are represented by Hopfield coefficients,[2] H_x and H_{cav}, respectively, which are shown in Fig. 1(b) as a function of the cavity detuning. In the present study, we employed two-dimensional coherent spectroscopy to investigate exciton-cavity coupling over a range of detuning values near the zero detuning of the cavity.

Two dimensional coherent spectroscopy (2DCS) is an enhanced version of transient four wave mixing (TFWM) which has been developed over the last two decades into an important tool for studying ultrafast coherent processes in semiconductor nanostructures.[3–8] Multidimensional spectroscopy has numerous advantages over linear

Send correspondence to Alan.Bristow@mail.wvu.edu

Paul, Jagannath; Wahlstrand, Jared; Bristow, Alan. "Polarization-dependent absolute-phase-corrected multidimensional coherent spectra of exciton-polaritons." Paper presented at Ultrafast Phenomena and Nanophotonics XXIV, San Francisco, CA, US. February 03, 2020 - February 06, 2020.

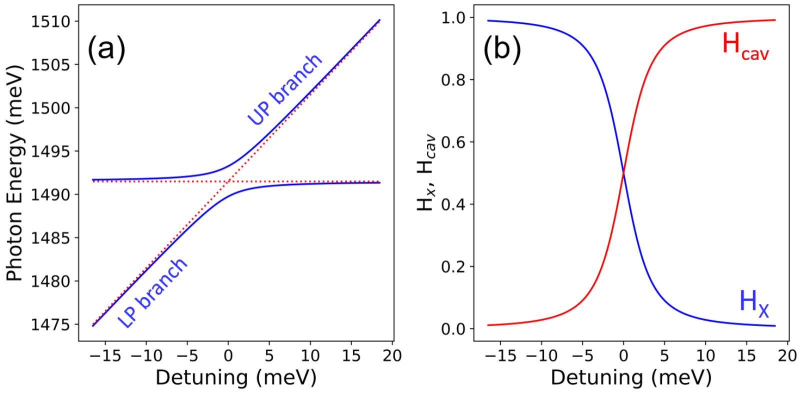


Figure 1. (a) Illustration of anticrossing of LP and UP branches in the semiconductor microcavity studied. The branch energies are shown as blue solid lines. The red dotted lines show the bare exciton (horizontal) and cavity mode (linear slope) energies in the absence of interaction. The splitting at zero detuning is the vacuum Rabi splitting E_{VRS}. (b) Hopfield coefficients of the UP branch plotted as a function of the cavity detuning. H_x and H_{cav} represent the exciton and photon fractions of the branch, such that $H_x + H_{cav} = 1$. H_x and H_{cav} are reversed for the LP branch.

spectroscopy techniques or conventional one dimensional time domain spectroscopy techniques as it correlates two time domains and represents the correlation spectra in a two dimensional map in the frequency domain. In a 2D spectrum, the resonant transition frequencies appear on the diagonal and any coupling between the resonances are revealed by the cross diagonal peaks. Another advantage of 2DCS is that homogeneous broadening can be easily isolated from inhomogeneous broadening a rephasing 2DCS scan. The homogeneous and inhomogeneous linewidths of a resonance can be directly obtained by finding cross-diagonal and diagonal slices of the resonance lineshape,[9] respectively, in a rephasing 2D spectrum. Polarization dependent 2DCS measurements isolate various quantum pathways and allow identification of many body interactions that contribute to the third order nonlinear response of the material.[10] Proper phasing of the 2D spectra allows extracting the signal phase. Previous work on quantum wells in microcavities reported magnitude 2D spectra, not fully phased spectra.[11–14]

In this proceeding, we present polarization-dependent complex 2D spectra, phased using spectrally-resolved transient absorption (SRTA) measurements. Real spectra of rephasing 2DCS for a range of cavity detuning values are measured as well as the polarization dependence for a single (near-zero) detuning. To our knowledge, this is the first report of phased 2DCS in microcavity samples and is the first step to understanding the full $\chi^{(3)}$ response of this strongly nonlinear optical system.

2. EXPERIMENTAL DETAILS

The experimental setup used in our study is well documented in Ref. 15 and will be only briefly described here. A 120 fs mode locked laser pulse from a Ti:sapphire oscillator with a repetition rate of 76 MHz is used as an input to a Multidimensional Optical Nonlinear SpecTrometeR (MONSTR)[16] instrument to generate four identical phase stabilized pulses. In Fig. 2(a), the pulse sequence for a rephasing scan is shown, where the phase conjugated pulse A* and pulse B are separated by time delay τ, and pulse B and pulse C are separated by time delay T. The beams are arranged in a box geometry, i.e. on the corner of a square, as shown in Fig. 2(b). The three excitation pulses, separated by two time delays τ and T, are focused on the sample mounted inside a liquid He cooled cryostat using a single lens and the resulting four-wave mixing (FWM) signal is emitted in the phase matching direction (along the fourth pulse). In a rephasing pulse sequence, the phase conjugated pulse arrives first on the sample followed by the other two pulses, and the FWM signal is emitted as a photon echo in the phase matched direction ($-k_A + k_B + k_C$, where k_i is the wavevector of excitation pulse i). The FWM signal is then collinearly combined with a phase stabilized local oscillator and the resulting spectral interferogram is dispersed into a grating spectrometer, and is then detected by a thermoelectrically cooled CCD. In this work,

Paul, Jagannath; Wahlstrand, Jared; Bristow, Alan. "Polarization-dependent absolute-phase-corrected multidimensional coherent spectra of exciton-polaritons." Paper presented at Ultrafast Phenomena and Nanophotonics XXIV, San Francisco, CA, US. February 03, 2020 - February 06, 2020.

rephasing 2D spectra were measured for four different polarization configurations and the time delay T is set to $\approx$ 50 fs. The polarizations of the incident laser pulses were controlled by liquid crystal variable retarders. In order to reduce artifacts caused by scattered light in the direction of the FWM signal, the phase of pump pulses A and B were phase cycled using two additional liquid crystal variable retarders. Phase cycling also improves the signal to noise ratio.

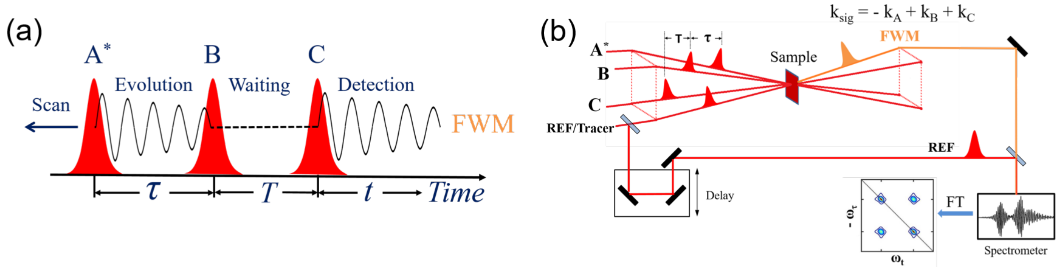


Figure 2. (a) Pulse sequence for a rephasing scan. (b) Pulse arrangement in the box geometry used in the experiment. The four-wave mixing (FWM) beam is emitted in a background-free, phase matching direction $-k_A + k_B + k_C$. The local oscillator (REF) is sent around the sample.

The microcavity sample used in our study was grown by molecular beam epitaxy on a GaAs substrate. The mirrors consist of GaAs/AlAs (14.5 and 12 bilayer) distributed Bragg reflectors separated by a wedged λ GaAs cavity, with a cavity mode close to 830 nm (1491 meV), the energy of the heavy hole exciton. A single 8 nm In$_{0.4}$Ga$_{0.96}$As QW is placed at the antinode of the cavity.[1,12] The sample has a vacuum Rabi splitting parameter $E_{\mathrm{VRS}} = 4$ meV. The cavity is slightly wedged, which allows us to vary the detuning by translating the sample in a direction perpendicular to the incident laser beams. Polaritons are very sensitive to the angle of incidence. In our experiment, the three excitation laser beams were incident on the sample at an external angle of approximately 7° to the normal.

Because of the strong interaction of the microcavity with the light, the optical density and dispersion of the sample is much higher than a typical semiconductor sample. As a result, the all-optical phasing technique often used with the MONSTR did not reliably work. Instead, we measure the spectrally-resolved transient absorption (SRTA) spectrum for the VVVV polarization configuration at the beginning of each scan. This spectrum is the same as a slice of the phased 2D spectrum, which allows us to fit it to find the global phase for that polarization.[17,18] Because we use liquid crystal variable retarders to polarize the beams and capture data for each polarization configuration on every step in the 2D scan, we can apply the known polarization-dependent phase imparted on the beams, which allows us to set the global phase of each polarization configuration measured.[15] Phased complex two-dimensional coherent spectra (2DCS) for common polarization configurations were recorded for a GaAs multiple quantum well sample for comparison to previous work.[10,15]

3. RESULTS AND DISCUSSION

Magnitude-only 2D spectra for this microcavity sample were presented in Refs. 12 and 14. In Ref. 15, preliminary phased polarization-dependent 2D spectra were presented for $\Delta = -6.0$ meV, where the lower branch is primarily light-like and the upper branch is primarily exciton-like. It was found that the diagonal peak corresponding to the exciton-like branch had the same dispersive lineshape as seen for the heavy hole exciton in a bare quantum well sample.[10] However, the sign of the biexciton feature was opposite what is seen in a bare quantum well sample. Here we present phased 2D spectra for values of detuning nearer the avoided crossing ($\Delta \approx 0$) and show how the phased 2D spectrum evolves with Δ.

Figure 3 depicts the rephasing amplitude (a-d) and real part (e-h) of the 2D spectrum for four different polarization configurations at near zero detuning ($\Delta \approx 0.05$ meV). Here, VVVV, VHHV, $\sigma^+\sigma^+\sigma^+\sigma^+$, and $\sigma^-\sigma^-\sigma^+\sigma^+$ represent the co-linear, cross-linear, co-circular, and cross-circular polarization configurations, respectively with first three letters representing the polarizations of the first three pulses, while the fourth letter corresponds to the emission polarization. The lower and upper polariton branch energies on the diagonal are labeled as LP and UP, respectively. Cross-peaks corresponding to couplings between the branches are labeled as UP-LP and LP-UP.

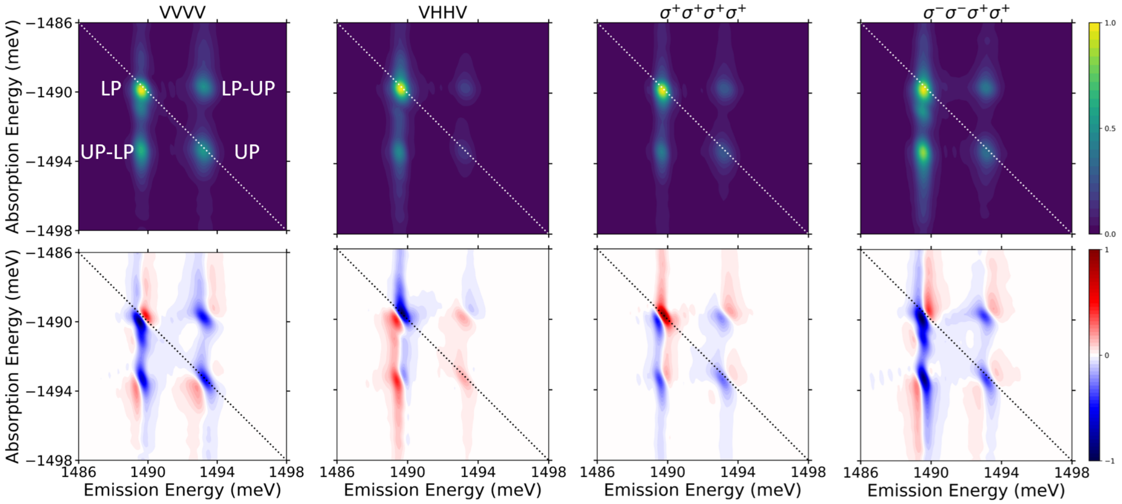


Figure 3. Polarization dependent rephasing 2D Spectra at near zero detuning, $\Delta \approx 0.05$ (meV). (a)-(h) The magnitude and the real part of the complex 2D spectra are shown in the top and bottom rows, respectively. VVVV, VHHV, $\sigma^+\sigma^+\sigma^+\sigma^+$, and $\sigma^-\sigma^-\sigma^+\sigma^+$ represent the co-linear, cross-linear, co-circular, and cross-circular polarization configurations, respectively.

The detuning dependence of the amplitude spectra we measure are consistent with previous results on this sample,[12] so we focus here on the real part spectra. Fig. 4 shows the real part of the rephasing spectrum for Δ values ranging from -3.27 meV to 3.37 meV.

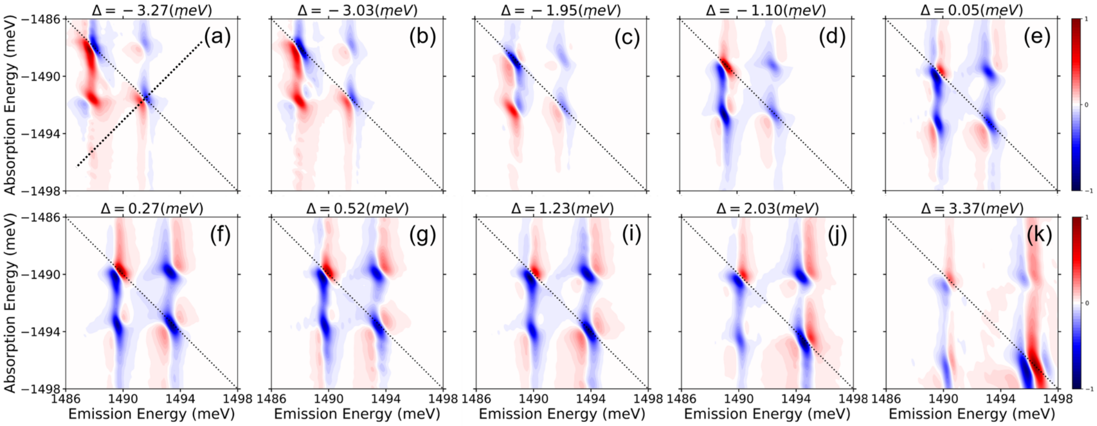


Figure 4. (a)-(j) Real Part of the rephasing 2D Spectra for different detuning values (Δ) in the range of -3.27 meV to 3.37 meV, (shown on top of each spectrum), for co-linear polarization configuration. Each spectrum is normalized to the maximum of the two diagonal resonances.

We find that the phase of the spectrum evolves approximately smoothly as we change the detuning parameter, as expected. This can be seen most clearly in Fig. 5, which shows line outs of the 2D spectra as a function of Δ.

A detailed analysis of the phased spectra and comparison to theoretical models will be explored in future work.

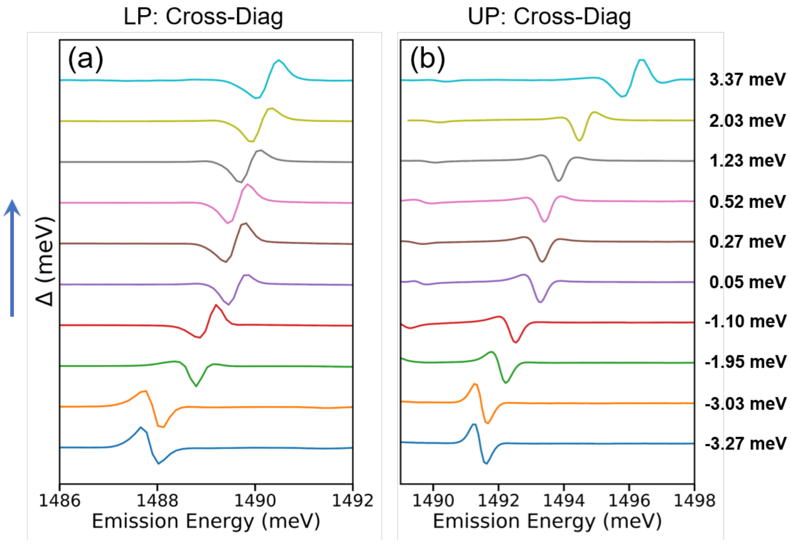


Figure 5. Cross-diagonal profile of real part 2D spectra of LP and UP as a function of detuning.

4. CONCLUSION

In summary, we have measured the response of a microcavity using multidimensional optical coherent spectroscopy. We investigate the coherent coupling between the exciton-like and cavity like normal modes in a semiconductor microcavity as a function of the detuning of the cavity. A full set of polarization dependent 2D spectra with co-linear, cross-linear, co-circular, and cross-circular polarization configuration were collected and phased with the help of spectrally-resolved transient absorption (SRTA) measurements. We observe the phase of the upper and lower polariton branches evolve gradually as a function of detuning.

ACKNOWLEDGMENTS

We thank G. Wernsing for technical assistance.

REFERENCES

[1] Khitrova, G., Gibbs, H. M., Jahnke, F., Kira, M., and Koch, S. W., “Nonlinear optics of normal-mode-coupling semiconductor microcavities,” *Rev. Mod. Phys.* **71**, 1591–1639 (Oct 1999).

[2] Deng, H., Haug, H., and Yamamoto, Y., “Exciton-polariton bose-einstein condensation,” *Rev. Mod. Phys.* **82**, 1489–1537 (May 2010).

[3] Dai, X., Richter, M., Li, H., Bristow, A. D., Falvo, C., Mukamel, S., and Cundiff, S. T., “Two-dimensional double-quantum spectra reveal collective resonances in an atomic vapor,” *Phys. Rev. Lett.* **108**, 193201 (May 2012).

[4] Moody, G., Siemens, M. E., Bristow, A. D., Dai, X., Karaiskaj, D., Bracker, A. S., Gammon, D., and Cundiff, S. T., “Exciton-exciton and exciton-phonon interactions in an interfacial gaas quantum dot ensemble,” *Phys. Rev. B* **83**, 115324 (Mar 2011).

[5] Paul, J., Stevens, C. E., Liu, C., Dey, P., McIntyre, C., Turkowski, V., Reno, J. L., Hilton, D. J., and Karaiskaj, D., “Strong quantum coherence between fermi liquid mahan excitons,” *Phys. Rev. Lett.* **116**, 157401 (Apr 2016).

[6] Paul, J., Stevens, C. E., Zhang, H., Dey, P., McGinty, D., McGill, S. A., Smith, R. P., Reno, J. L., Turkowski, V., Perakis, I. E., Hilton, D. J., and Karaiskaj, D., “Coulomb-interaction induced coupling of landau levels in intrinsic and modulation-doped quantum wells,” *Phys. Rev. B* **95**, 245314 (Jun 2017).

[7] Dey, P., Paul, J., Glikin, N., Kovalyuk, Z. D., Kudrynskyi, Z. R., Romero, A. H., and Karaiskaj, D., “Mechanism of excitonic dephasing in layered inse crystals,” *Phys. Rev. B* **89**, 125128 (Mar 2014).

[8] Dey, P., Paul, J., Moody, G., Stevens, C. E., Glikin, N., Kovalyuk, Z. D., Kudrynskyi, Z. R., Romero, A. H., Cantarero, A., Hilton, D. J., and Karaiskaj, D., “Biexciton formation and exciton coherent coupling in layered gase,” *The Journal of Chemical Physics* **142**(21), 212422 (2015).

[9] Siemens, M. E., Moody, G., Li, H., Bristow, A. D., and Cundiff, S. T., “Resonance lineshapes in two-dimensional fourier transform spectroscopy,” *Opt. Express* **18**, 17699–17708 (Aug 2010).

[10] Bristow, A. D., Karaiskaj, D., Dai, X., Mirin, R. P., and Cundiff, S. T., “Polarization dependence of semiconductor exciton and biexciton contributions to phase-resolved optical two-dimensional Fourier-transform spectra,” *Physical Review B* **79**, 161305 (Apr. 2009).

[11] Wen, P., Christmann, G., Baumberg, J. J., and Nelson, K. A., “Influence of multi-exciton correlations on nonlinear polariton dynamics in semiconductor microcavities,” *New Journal of Physics* **15**, 025005 (feb 2013).

[12] Wilmer, B. L., Passmann, F., Gehl, M., Khitrova, G., and Bristow, A. D., “Multidimensional coherent spectroscopy of a semiconductor microcavity,” *Phys. Rev. B* **91**, 201304 (May 2015).

[13] Takemura, N., Trebaol, S., Anderson, M. D., Kohnle, V., Léger, Y., Oberli, D. Y., Portella-Oberli, M. T., and Deveaud, B., “Two-dimensional Fourier transform spectroscopy of exciton-polaritons and their interactions,” *Physical Review B* **92**, 125415 (Sept. 2015).

[14] B. Wilmer, F. Passmann, M. G. G. K. and Bristow, A. D., “Multidimensional coherent spectroscopy of a semiconductor microcavity,” *Proc. SPIE* **9746**, 97461B (2016).

[15] Wahlstrand, J. K., Wernsing, G. M., Paul, J., and Bristow, A. D., “Automated polarization-dependent multidimensional coherent spectroscopy phased using transient absorption,” *Opt. Express* **27**, 31790–31799 (Oct 2019).

[16] Bristow, A. D., Karaiskaj, D., Dai, X., Zhang, T., Carlsson, C., Hagen, K. R., Jimenez, R., and Cundiff, S. T., “A versatile ultrastable platform for optical multidimensional fourier-transform spectroscopy,” *Review of Scientific Instruments* **80**(7), 073108 (2009).

[17] Zhang, T., Borca, C. N., Li, X., and Cundiff, S. T., “Optical two-dimensional fourier transform spectroscopy with active interferometric stabilization,” *Opt. Express* **13**, 7432–7441 (Sep 2005).

[18] Gallagher Faeder, S. M. and Jonas, D. M., “Two-dimensional electronic correlation and relaxation spectra: theory and model calculations,” *The Journal of Physical Chemistry A* **103**(49), 10489–10505 (1999).

A Suite of Metrics for Calculating the Most Significant Security Relevant Software Flaw Types

Peter Mell
National Institute of Standards and Technology
Gaithersburg MD, USA
peter.mell@nist.gov

Assane Gueye
University Alioune Diop, Bambey-Senegal
Prometheus Computing, LLC
Bambey, Senegal
a.gueye@prometheuscomputing.com

***Abstract*—The Common Weakness Enumeration (CWE) is a prominent list of software weakness types. This list is used by vulnerability databases to describe the underlying security flaws within analyzed vulnerabilities. This linkage opens the possibility of using the analysis of software vulnerabilities to identify the most significant weaknesses that enable those vulnerabilities. We accomplish this through creating mashup views combining CWE weakness taxonomies with vulnerability analysis data. The resulting graphs have CWEs as nodes, edges derived from multiple CWE taxonomies, and nodes adorned with vulnerability analysis information (propagated from children to parents). Using these graphs, we develop a suite of metrics to identify the most significant weakness types (using the perspectives of frequency, impact, exploitability, and overall severity).**

***Index Terms*—Metrics, Software flaws, Vulnerabilities**

I. Introduction

The Common Weakness Enumeration (CWE) [1] [2] is a prominent list of software weakness types. It is maintained by the MITRE corporation, funded by the United States (U.S.) government, and developed with the participation of 55 organizations. The CWE list contains 808 weaknesses organized by multiple views. Views are 'hierarchical representations' of CWEs (i.e., taxonomies) serving different communities with different perspectives on the data.

A specific view, 1003, was created to support the labelling of publicly disclosed software vulnerabilities ('potential weaknesses within sources that handle public, third-party vulnerability information'). It contains 123 software flaws. The National Vulnerability Database (NVD) [3] and other vulnerability databases and security tools use view 1003 to describe the underlying security flaws within analyzed vulnerabilities. 98.8 % of the 12 760 fully analyzed vulnerabilities published by NVD in 2019 were able to be mapped to view 1003, demonstrating its applicability and coverage.

This linkage of vulnerability analysis to the view 1003 CWEs opens the possibility of using the NVD analysis of software vulnerabilities to identify the most significant weaknesses that enable those vulnerabilities. In this work, we accomplish this through creating mashup views combining the following resources:

- the multiple primary taxonomies of the CWE (views 1003, 1000, 699, and 1008),
- the Common Vulnerabilities and Exposures (CVE) [4] enumeration of publicly disclosed software vulnerabilities,
- the NVD mapping of CVEs to view 1003 CWEs, and
- the NVD measurements of each CVE using the Common Vulnerability Scoring System (CVSS) [5] [6]. This calculates the exploitability, impact, and overall severity of each CVE outside of any particular deployment context.

The result of creating mashups of this data are graphs that have CWEs as nodes. The edges between the nodes are extracted from the parent-child relationships between the multiple CWE taxonomies. And the nodes are labelled with CVE and CVSS information (propagated backwards along the edges). We apply to these graphs a suite of simple metrics that we developed to identify the most 'significant' weakness types. We evaluate significance from multiple perspectives using metrics focused on the following areas: frequency, impact, and exploitability. In doing this we evaluate the CWEs in two distinct groups to take into account the varying levels of abstraction of the CWEs.

We create most significant weakness lists for each metric for the CVE vulnerabilities published in 2019 (not provided due to space limitations). We then analyze the differences between these six lists (3 metrics * 2 sets of CWE types) using two algorithms for comparing differences between ordered lists (Kendall's Tau and the Spearman's footrule variant [7]). We find that different weaknesses tend to emerge as the most significant depending upon the perspective, the metric used, and the CWE type. Note that we use simple low level metrics for our perspectives. This is because there is no ground truth for aggregating those metrics; equations in security that aggregate simple metrics are often practically useful but less scientifically defensible.

Finally, we note that the CWE already has an official metric to identify the 'most dangerous' CWEs. It aggregates both frequency and severity with severity itself being an aggregate metric combining exploitability and impact. We discover weaknesses with this official metric that leads to the under counting of certain CWEs.

We recommend that software developers and creators of software bug finding tools use our approach to prioritize finding and eliminating these most significant weaknesses to

Mell, Peter; Gueye, Assane. "A Suite of Metrics for Calculating the Most Significant Security Relevant Software Flaw Types." Paper presented at 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, ES. July 13, 2020 - July 17, 2020.

reduce the number and severity of security related flaws in software.

II. Background

As mashup research, our approach combines multiple resources. These are briefly described and referenced here.

A. Common Weakness Enumeration

Our research is primarily focused on the Common Weakness Enumeration (CWE) [8], a 'community-developed list of common software security weaknesses'. 'It serves as a common language, a measuring stick for software security tools, and as a baseline for weakness identification, mitigation, and prevention efforts' [9]. The 808 software weaknesses within the enumeration are referred to as CWEs where each is named CWE-X with X being some integer. Each CWE is characterized as either a class, base, variant, or compound. Classes are the highest level of abstraction, followed by bases, and then by variants. Compounds are relatively rare and are combinations of multiple bases and/or variants. In our work we evaluate classes separately from bases, variants, and compounds, given that the classes have a much higher level of abstraction.

Besides the CWE weaknesses, there are also 295 categories and 38 views. Confusingly, these are also considered CWEs; for simplification we use the name CWE to refer only to the weakness CWEs. The categories are used to organize the CWEs within select views (this is not used in our research). The views are hierarchical organizations of a subset of CWEs according to some perspective (essentially a taxonomy). The three primary taxonomies are the 'Research Concepts' (view 1000), 'Development Concepts' (view 699), and 'Architectural Concepts' (view 1008). This last view, 1008, was not useful to our work and is not used because it doesn't provide a hierarchy of CWEs but instead uses the categories to group CWEs. The view 1003 designed for vulnerability databases, mentioned previously, is called 'CWE Weaknesses for Simplified Mapping of Published Vulnerabilities View' and is the core data structure upon which our work builds.

B. Common Vulnerabilities and Exposures

The set of software vulnerabilities used for this research come from the Common Vulnerabilities and Exposures (CVE) program, maintained by the MITRE corporation. 'CVE is a list of entries—each containing an identification number, a description, and at least one public reference—for publicly known cybersecurity vulnerabilities' [4] [10].

C. Common Vulnerability Scoring System

The Common Vulnerability Scoring System 'provides a way to capture the principal characteristics of a vulnerability and produce a numerical score reflecting its severity' [11]. It provides equations for calculating a vulnerability's base score (inherent risk outside of any particular environment), temporal score (changing risk over time), and environmental score (risk within a particular environment). We use the base score, which is composed of two sub-scores that calculate the exploitability and impact of a vulnerability. It is maintained by the Forum of Incident Response and Security Teams (FIRST). The detailed specification for CVSS version 3.1 is available at [5].

D. National Vulnerability Database

The National Vulnerability Database (NVD) is 'the U.S. government repository of standards based vulnerability management data' [3]. It is maintained by the U.S. National Institute of Standards and Technology. We use its scoring of CVEs with CVSS scores and its mapping of the CVEs to view 1003 CWEs.

III. Foundational Data Structures

This section describes how we generate the foundational data structure used by our metrics to calculate the most significant security relevant software flaw types. We generate a directed acyclic graph (DAG) of CWEs that we will use to propagate CWE analysis data between the CWEs.

A. View 1003 Graph

We begin with the set of CWEs in CWE view 1003 since that is the set that was adopted by the NVD (and is the set identified by MITRE as most applicable to CVE vulnerabilities). We then form a graph of the view 1003 nodes through extracting the 'ChildOf' relationships in the CWE view 1003 Extensible Markup Language (XML) file. Other kinds of relationships are provided in the XML file but we don't use them because none of them definitively indicate the parent child relationship needed to construct edges in our graph (for example, 'CanPrecede'). The result is a rooted tree[1] with the root being CWE 1003, the nodes at distance one from the root being classes, and the nodes at distance 2 being bases, variants, and compounds. We remove the root as we are only interested in the classes, bases, variants, and compounds. The resulting DAG has 123 nodes and 87 edges, shown in Figure 1. On the left side are the 36 class nodes in blue. The majority of class nodes have edges to bases, variants, or compounds, but five do not. On the right side, the largest grouping of nodes in a single column in purple represents the 82 bases. Moved slightly to the right and in green are the 3 variants. Moved even farther to the right in orange are the 2 compounds.

B. Direct Edge Augmentation

We next augment our view 1003 DAG with edges extracted from the 'ChildOf' relationships specified within other CWE view XML files. For this we use both the CWE research and development concepts views (essentially alternate taxonomies). We can do this because for our metrics we aren't focused on a particular type of child-parent relationship, we just want to know that a child-parent relationship definitively exists between some pair of CWEs in the view 1003 set. This analysis adds 19 edges, shown in green in Figure 2. Note that we move three of the class nodes slightly left of the main

[1] A perfect tree structure is uncommon in weakness/vulnerability taxonomies. This encouraged us to explore possible missing relationships.

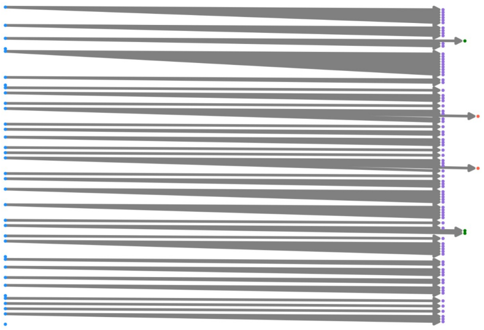


Fig. 1. CWE View 1003 (123 nodes, 87 edges)

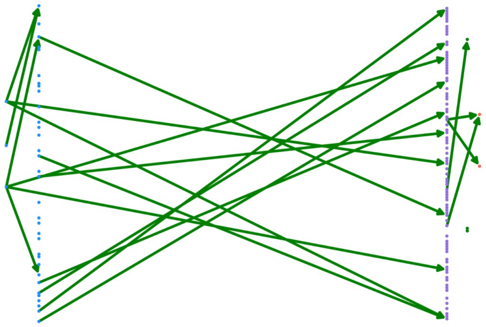


Fig. 2. CWE View 1003 Nodes with Direct Edges from Non-1003 Views (123 nodes, 19 edges)

column of class nodes to enhance visibility because they now have edges to other classes.

C. Indirect Edge Augmentation

Lastly, we create a new DAG (to be used temporarily for this section's analysis) by unifying the set of nodes in views 1003, 1000, and 699 and then adding edges based on the 'ChildOf' relationships specified in the three XML view files. This produces a DAG with 834 CWEs and 1046 edges. Then for each pair of nodes within view 1003, we determine if a path exists connecting them that uses at least one node not in view 1003. Each such discovered path can be used to add an edge to our foundational data structure DAG. These 29 'indirect' edges (that really represent paths using nodes not shown) can be seen in blue in Figure 3.

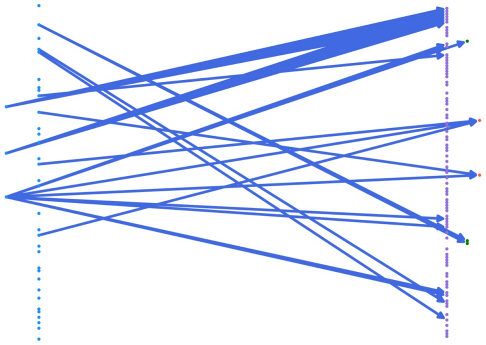


Fig. 3. CWE View 1003 Nodes with Edges Representing Paths from Non-1003 Views (123 nodes, 29 edges)

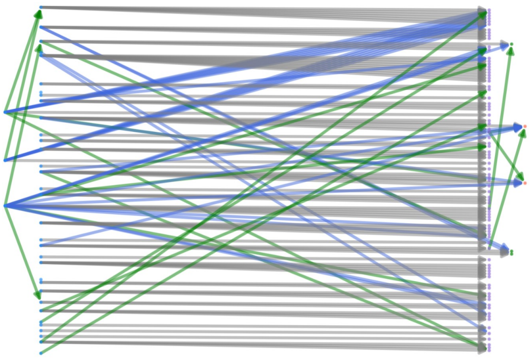


Fig. 4. Composite Graph of View 1003 with Direct Edges and Edges Representing Paths from Non-1003 Views (123 nodes, 135 edges)

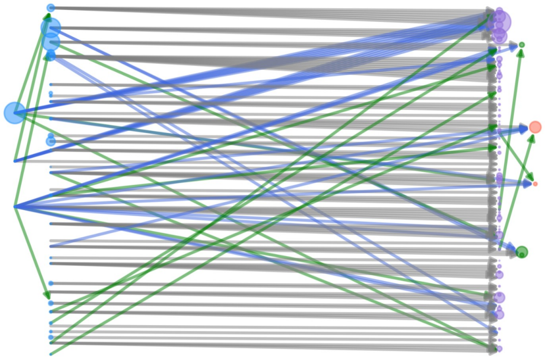


Fig. 5. View 1003 Nodes Adorned with NVD Data (no propagation)

D. Composite Directed Acyclic Graph

We now put together our DAG representing the 1003 view with the direct edge augmentation from Section III-B and the indirect edge augmentation from Section III-C. The resulting graph is shown in Figure 4. It has 123 nodes and 135 edges.

E. Node Adornment

The next step is to adorn the DAG with vulnerability analysis data from the NVD. We take each CVE in NVD that has one or more CWE mappings, and we label each relevant CWE node in the DAG with a vector containing the CVE name, the publish date, and the CVSS attribute information. Figure 5 shows this adornment for the CVEs published in 2019. Note that the size of each node now represents the number of vulnerability vectors mapped to that node.

F. Data Propagation

The edges within the DAG represent opportunities for propagating vector data between CWEs. Parent CWEs receive

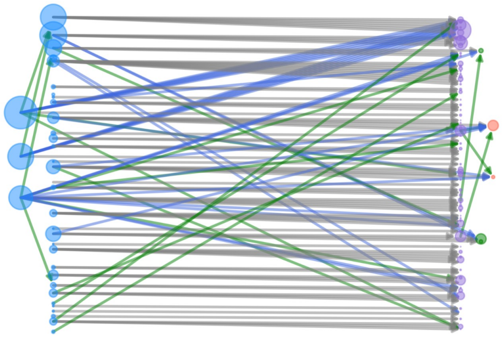


Fig. 6. View 1003 Nodes Adorned with NVD Data (with propagation)

the vectors of their children (with any duplicates being removed). This is because if a vector applies to a CWE then it by definition applies to its more general parent. Also, we discovered that NVD analysts only label a CVE with its most specific CWE. They do not label a CVE with a class if they can determine the applicable base, variant, or compound within that class. Figure 6 shows the DAG adorned with the 2019 vulnerability vectors propagated from children to parents. Note, by comparing figures 5 and 6, how without the propagation some classes get under counted (especially those with many popular bases that are the class' children).

IV. Metrics for Calculating Significance

The DAG in Figure 6 is what we use as the foundational data structure to calculate three simple metrics. The metrics we calculate on this DAG are normalized frequency, mean exploitability, and mean impact. These three metrics are defined below.

We start with a metric to count the number of CVEs mapped to each CWE. Let I designate the set of all CWEs and let J be the set of all CVEs. For CWE $i \in I$, let N_i be the number of CVEs mapped to i. We can write it as:

$$N_i = \sum_{j \in J} e_{ij}, \tag{1}$$

where

$$e_{ij} = \begin{cases} 1, & \text{if CVE } j \text{ is mapped to CWE } i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

A. Metric for normalized frequency (F_i)

$$F_i = \frac{N_i - \min\limits_{i' \in I}(N_{i'})}{\max\limits_{i' \in I}(N_{i'}) - \min\limits_{i' \in I}(N_{i'})}. \tag{3}$$

B. Metric for mean exploitability (Q_i)

Let q_j be the CVSS exploitability score for CVE j. We can write the average of the q_j in all CVEs mapped to CWE i as:

$$\overline{Q_i} = \frac{\sum_{j \in J} q_j e_{ij}}{N_i}. \tag{4}$$

C. Metric for mean impact (R_i)

Let r_j be the CVSS impact score CVE j. We can write the average of the r_j in all CVEs mapped to CWE i as:

$$\overline{R_i} = \frac{\sum_{j \in J} r_j e_{ij}}{N_i}. \tag{5}$$

V. Weaknesses in the Official CWE Equation

In September 2019, the official CWE website provided a metric for measuring the 'CWE Top 25 Most Dangerous Software Errors' [12]. It is an aggregate metric combining the normalized frequency of CVEs mapped to CWEs while using the CVSS severity calculated for each mapped CVE.

This metric, like ours, combines together CWE and NVD data, evaluating only the CWEs within the 1003 view. Differing from ours, it uses the raw NVD mappings (it doesn't perform any data propagation) and evaluates all CWE types together (i.e., classes, bases, variants, and compounds). The metric is described in [12], we summarize it below (leveraging two of our equations from section IV).

A. Official CWE Metric

We first need to define the mean CVSS score for some CWE i as $\overline{S_i}$. Let s_j be the CVSS base score for CVE j. We can write the average of the s_j in all CVEs mapped to CWE i as:

$$\overline{S_i} = \frac{\sum_{j \in J} s_j e_{ij}}{N_i}. \tag{6}$$

Now we define the official 'most dangerous' CWE score as D_i for some CWE i. Let F_i refer to equation 3 and let c_j be the CVSS score for the j-th CVE.

$$D_i = F_i * \frac{\overline{S_i} - \min\limits_{j \in J}(c_j)}{\max\limits_{j \in J}(c_j) - \min\limits_{j \in J}(c_j)} * 100. \tag{7}$$

B. Weakness 1: Undercounting Parent CWEs

Almost all CWE classes in view 1003 have children, as do a few bases. All CWEs that are parents then get under counted because CVEs that apply to them are often assigned to their children. The official CWE metric does not propagate CVE assignments from children to parents. Also, the NVD analysts assign only the most specific CWE to a CVE; they do not include the parents of marked CWEs. This artificially decreases the importance of CWEs that have children when using the official metric. Using the official CWE metric on the 2019 data, each CWE gets assigned a mean of 87.99 CVEs. Using the propagation proposed in this paper to avoid under counting parents, each CWE gets assigned a mean of 294.71 CVEs.

C. Weakness 2: Class Bias

The inclusion of all CWE types in the official metric (i.e., classes, bases, variants, and compounds) causes some classes to be unfairly promoted to being within the top lists. This is because classes are at a much higher level of abstraction and thus more CVEs will apply to them. Using the official metric, we find that there are 8 classes on the 2019 top 25 list (32 %) while there are 36 classes out of 123 CWEs in the 1003 view (29 %). While a bias is not particularly apparent here, the bias is muted because the classes are under counted due to weakness 1 (above). To correctly isolate and measure weakness 2, we remove weakness 1 by using the official metric but while performing data propagation from children to parents. Our regenerated 2019 official top 25 list then contains 16 classes (64 %); here the classes are vastly over represented since only 29 % of the CWEs in view 1003 are classes.

VI. Analysis

Our approach of propagating CVE data over the CWE taxonomies fills in data missing from the official CWE metric approach. This addresses weakness 1 described in Section V-B. Our approach a creating separate top lists for the two levels of CWE abstraction addresses weakness 2, described in Section V-C. Thus, we argue that our approach improves over the original. The question though is whether these improvements make any difference in the results.

Using our DAG and the three metrics, we calculated the most significant CWEs for 2019 at the two levels of abstraction. We now evaluate these results to verify that propagating analysis data over our DAG substantially changes the generated most significant software flaw lists. We also verify that our multiple metrics produce substantially different lists. To compare different rankings, we measure their distance using two related metrics: the Kendall's Tau and the Spearman's Footrule [7]. For two rankings l_1 and l_2, the Kendall's Tau $K(l_1, l_2)$ measures the number of pairs of elements in l_1 that are swapped in their relative positions in l_2. The Spearman's Footrule $F(l_1, l_2)$ measures the number of adjacent element swaps that would need to be performed in l_1 to convert it into l_2. It has been proven that $\forall l_1, l_2, K(l_1, l_2) \leq F(l_1, l_2) \leq 2K(l_1, l_2)$ [7].

Both approaches require that the rankings be of the same length and contain the same elements. Thus, when comparing rankings we use the full rankings of all CWEs observed in the data as opposed to comparing top X lists where X is some integer (using some X to limit list length usually results in lists that contain at least one distinct CWE). The number of observed class CWEs in our data was 36 and the number of non-class CWEs was 87. We performed an empirical study to determine K() and F() for random lists of these sizes using 100 000 trials. The results are shown in Table I.

TABLE I
Distance Between Random Rankings

	K()	F()
Length 36	315	432
Length 87	1870	2522

We first verify that propagating data over our DAG substantially changes the rankings. For each of our metrics, we calculate the full ranking using all available CWEs and then compare that against a ranking created using the same metric but without propagating data over our DAG. The results are shown in Table II. Overall, they show that the lists do change significantly when using the DAG to propagate data. Note that the distances are less for the non-class CWEs which is remarkable because those lists are more than twice as long as the class CWE lists (longer lists in general produce greater distances due to more elements possibly being out of place). However, this can be explained by noting that there are only 4 edges between the non-class CWEs which diminishes the effect of propagating data from children to parents. There are, on the other hand, 135 edges over which data can be propagated to the class CWEs.

We next verify that our multiple metrics produce substantially different lists. If this were not the case, then that would argue towards producing just a single list as opposed to multiple lists with different perspectives. Table III shows the results for the class CWE lists and Table IV shows the results for the non-class CWE lists. Overall, all the lists appear different. The mean exploitability lists have the most distinction from the other two. Comparing the mean exploitability and normalized frequencies lists for the class CWEs we get lists that are even more different than random (see Table I). Comparing the mean exploitability and the mean impact lists, they are almost as different as the random lists.

VII. Related Work

The NVD also provides CWE rating data. This is in the form of two visualizations that show the relative frequency between the observed CWEs per year and another that shows the actual frequency change for the most frequent CWEs [13]. This data is incorrect because it miscounts the frequencies of the CWEs that have children in the CWE taxonomies (because NVD only labels a CVE with its most specific CWE and this information is not propagated to its parents). Also, NVD doesn't distinguish between classes and bases/variants/compounds. The classes are larger categories biasing them to be very frequent, crowding out the bases that are less likely simply due to them being more specific.

VIII. Conclusion

The multiple CWE views can be evaluated as hierarchical taxonomies that reveal parent-child relationships between pairs of CWEs. The different perspectives for each view does not invalidate unifying them because in scoring a CWE as to its significance we want to know all applicable CVEs regardless of the particular method used to organize the CWEs hierarchically. View 1003 provides an obvious base taxonomy from which to start as it was designed to cover the CWEs most used by CVEs. However, its perfect tree structure indicates likely missing relationships. We find those relationships through evaluating the primary three CWE taxonomies (one of which we have to discard because it uses non-CWEs for its higher level classifications). We first find direct missing edges and then find indirect edges (those that represent paths traversing non-1003 view CWEs).

TABLE II
DISTANCE BETWEEN TOP LISTS CREATED USING RAW CWES VERSUS PROPAGATING DATA OVER THE CONSTRUCTED DAG

	Class CWEs		Non-Class CWEs	
	(list length of 36)		(list length of 87)	
	K()	F()	K()	F()
Normalized Frequency	213	324	115	228
Mean Exploitability	188	264	81	154
Mean Impact	175	256	82	164

TABLE III
DISTANCE BETWEEN TOP CLASS CWE LISTS CREATED USING DIFFERENT METRICS (WITH PROPAGATING DATA ON THE DAG)

	Normalized Frequency		Mean Exploitability		Mean Impact	
	Kendall Tau	Spearman	Kendall Tau	Spearman	Kendall Tau	Spearman
Normalized Frequency	0	0	357	488	213	290
Mean Exploitability			0	0	312	420
Mean Impact					0	0

TABLE IV
DISTANCE BETWEEN TOP NON-CLASS LISTS CREATED USING DIFFERENT METRICS (WITH PROPAGATING DATA ON THE DAG)

	Normalized Frequency		Mean Exploitability		Mean Impact	
	Kendall Tau	Spearman	Kendall Tau	Spearman	Kendall Tau	Spearman
Normalized Frequency	0	0	1281	1716	976	1350
Mean Exploitability			0	0	1527	1984
Mean Impact					0	0

The NVD is an ideal data source to analyze the CWEs because it both maps the CVEs to CWEs and also provides CVSS scores for each CVE. We adorned our unified DAG with the CVE information from NVD and propagated that information from children to parents. We then evaluated the DAG with three simple metrics; focusing separately on classes and bases/variants/compounds due to their very different levels of abstraction. We then generated top lists that provide the most significant CWEs relative to a particular perspective and abstraction. We analyzed those lists and discovered significant differences between them. This argues for the usefulness of and need for multiple top lists with different perspectives.

It is our hope that software developers and creators of software bug finding tools will use our approach to help prioritize finding and eliminating CWEs in their code. We hope in turn that this will help reduce the number and severity of security related flaws in software.

ACKNOWLEDGEMENT

This work was partially accomplished under NIST Cooperative Agreement No.70NANB19H063 with Prometheus Computing, LLC.

REFERENCES

[1] Y. Wu, I. Bojanova, and Y. Yesha, "They know your weaknesses–do you?: Reintroducing common weakness enumeration," *CrossTalk*, vol. 45, 2015.

[2] R. Martin, S. Barnum, and S. Christey, "Being explicit about security weaknesses," *Blackhat DC*, 2007.

[3] "National vulnerability database," 2019, accessed: 2019-12-10. [Online]. Available: https://https://nvd.nist.gov

[4] "Common vulnerabilities and exposures," 2019, accessed: 2019-12-10. [Online]. Available: https://cve.mitre.org

[5] "Common vulnerability scoring system version 3.1 specification document revision 1," 2019, accessed: 2019-12-10. [Online]. Available: https://www.first.org/cvss/v3.1/specification-document

[6] P. Mell, K. Scarfone, and S. Romanosky, "Common vulnerability scoring system," *IEEE Security & Privacy*, vol. 4, no. 6, pp. 85–89, 2006.

[7] "Generalized distances between rankings," 2010, accessed: 2019-12-10. [Online]. Available: https://tinyurl.com/theory-stanford-edu-sergei

[8] R. A. Martin and S. Barnum, "Common weakness enumeration (cwe) status update," *ACM SIGAda Ada Letters*, vol. 28, no. 1, pp. 88–91, 2008.

[9] "Common weakness enumeration," 2019, accessed: 2019-12-10. [Online]. Available: https://cwe.mitre.org

[10] D. W. Baker, S. M. Christey, W. H. Hill, and D. E. Mann, "The development of a common enumeration of vulnerabilities and exposures," in *Recent Advances in Intrusion Detection*, vol. 7, 1999, p. 9.

[11] "Common vulnerability scoring system special interest group," 2019, accessed: 2019-12-10. [Online]. Available: https://www.first.org/cvss

[12] "2019 cwe top 25 most dangerous software errors," 2019, accessed: 2019-12-10. [Online]. Available: https://cwe.mitre.org/top25/archive/2019/2019_cwe_top25.html

[13] "National vulnerability database, cwe over time," 2019, accessed: 2019-12-10. [Online]. Available: https://nvd.nist.gov/general/visualizations/vulnerability-visualizations/cwe-over-time

MACHINE LEARNING METHODS FOR PREDICTING SEISMIC RETROFIT COSTS

J.F. Fung(1), S. Sattar(2), D.T. Butry(3), S.L. McCabe(4)

(1) Economist, National Institute of Standards and Technology, juan.fung@nist.gov
(2) Research Structural Engineer, National Institute of Standards and Technology, siamak.sattar@nist.gov
(3) Economist, National Institute of Standards and Technology, david.butry@nist.gov
(4) NEHRP Director, National Institute of Standards and Technology, steven.mccabe@nist.gov

Abstract

Aging building clusters worldwide, especially in high seismic regions, will require a retrofit approach to improve the resilience of the built environment. One of the main challenges of retrofitting existing buildings is the associated cost. Fung et al. [1] develop a predictive modeling approach to estimating seismic retrofit costs. The predictive modeling approach uses historical data to predict retrofit costs for existing buildings based on observable building characteristics ("the features"). The advantages of this approach are that it is (1) cheap, (2) fast, and (3) can be applied to a single building or a large inventory of buildings. However, the approach relies on a linear model for prediction, which assumes a restrictive relationship between retrofit cost and the features. In this paper, we consider machine learning methods that can capture more complex, potentially nonlinear relationships between retrofit cost and the features. The paper considers ensemble methods (bagging and boosting) as well as neural networks and compares their performance to that of the linear model. The results show that a neural network provides the best performance in terms of prediction error. In applications, a user faces a tradeoff between accuracy and interpretability: while ensemble methods and deep neural networks may improve accuracy, they are not as easily interpretable as the linear model.

Keywords: seismic retrofit, risk reduction, resilience, retrofit cost estimation, building portfolio, prediction

Fung, Juan; Sattar, Siamak; Butry, David; McCabe, Steven. "Machine Learning Methods for Predicting Seismic Retrofit Costs." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Aging building clusters around the world, especially in high-seismicity regions, will require a retrofit approach to improve the resilience of the built environment. One of the main challenges of retrofitting existing buildings is the associated cost. The reliable estimation of retrofit cost is valuable for decision makers, especially in the planning stage of a potential construction project. Machine learning methods have the potential to accurately predict retrofit cost for a single building and, more importantly, for a large inventory of buildings.

Fung et al. [1] present a predictive modeling approach that predicts retrofit cost based on observable building characteristics (e.g., building age and size) and the target performance objective. However, the approach relies on a linear model that essentially assumes a linear relationship between the predictors and retrofit cost (in particular, a *generalized linear model*, or GLM). In this paper, we consider machine learning methods that can capture more complex and possibly nonlinear relationships (in particular, ensemble methods and neural networks). Such methods are useful only if they can perform better than the linear model. However, the tradeoff is a lack of interpretability: for more complex methods, it is often difficult to understand how the model arrives at a prediction. The paper considers whether the performance gain is sufficient to warrant lack of interpretability.

The use of machine learning methods in the retrofit cost prediction literature is not new. Jafarzadeh et al. [2] use linear regression and Jafarzadeh et al. [3] use artificial neural networks. However, a comparison between methodologies is not made. Nasrazadani et al. [4] use Bayesian linear regression. However, the main objective is not to accurately predict cost but rather to quantify uncertainty in covariate effects (e.g., lateral strength). In the broader context of construction cost prediction, Elfaki et al. [5] provide a literature review of "intelligent methods," including machine learning methods.

The paper is organized as follows. Section 2 presents a discussion of the machine learning methods used in this paper. The main results comparing performance of machine learning methods is presented in Sec. 3. Finally, Sec. 4 presents concluding remarks and directions for future research.

2. Machine Learning Methods

This section describes the machine learning methods used in this paper to predict seismic retrofit costs. It is worth noting that the methods covered are not intended to be exhaustive. Rather, the methods are chosen in order to illustrate the tradeoff between model complexity and performance. A more complete treatment of machine learning is beyond the scope of this paper. An excellent reference is Mitchell [6].

At a high level, machine learning is a set of methods that finds patterns in data [7]. There are two main types of machine learning methods: supervised learning, which is *predictive*, and unsupervised learning, which is *descriptive* [7]. The key difference is that supervised learning has access to an outcome of interest (e.g., retrofit cost) and the goal is to predict the outcome from the input data (e.g., building characteristics) [8]. In contrast, there is no explicit outcome of interest under unsupervised learning and the goal is to explore patterns in the input data (e.g., clusters or latent features) [8].

This paper deals with a supervised learning problem. In particular, retrofit cost prediction is a *regression* problem because the outcome of interest, retrofit cost, is real-valued. If the outcome of interest is categorical (e.g., high cost vs low cost), the supervised learning problem is *classification*. In supervised learning, the model learns about the relationship between the input data (or features) and the outcome of interest. See Hastie et al. [8] for a treatment of supervised learning problems.

More precisely, suppose you have a set of features X (the covariates, or explanatory variables) associated with an outcome Y (the dependent, or response, variable). There is some relationship between X and Y, say $Y = f(X)$, and you would like to "learn" that relationship in order to predict new values of the outcome when presented with new instances of the features, say, $Y^*_{new} = f(X_{new})$.

In practical terms, to learn the function f means to fit a model (e.g., a GLM or a neural network) to data. How well the model captures the relationship between features and outcome depends on two key factors. The first is that the "training data" used to fit the model, $\{(X_i, Y_i)\}_{i=1}^{n}$, is sufficiently rich to credibly estimate the parameters of the underlying model. This means that a large number of examples n is not sufficient; the quality of the training data also matters. The second is the model itself, which may require "tuning" (i.e., the choice of hyperparameters such as the number of layers in a neural network). These issues, which are beyond the scope of this paper, are covered extensively in Hastie et al. [8].

In the context of cost prediction, X includes building characteristics (such as building age and size), and retrofit characteristics (in particular, the performance objective). The outcome Y may represent the actual cost from a retrofit or the estimated cost (e.g., a design or bid estimate) for a potential retrofit. As in Fung et al. [1], this paper uses data collected for FEMA 156 [9]. The training data is discussed in more detail in Sec. 3 below.

To determine how well a model captures the relationship between features and outcome, we need to quantify the quality of the predictions the model produces. This requires the choice of a loss function, $L(Y, f(X))$, that penalizes errors in prediction [8]. In other words, the criterion for choosing f should be to *minimize prediction error*. In regression problems, the most common and convenient way to quantify prediction error is squared error loss (or mean squared error):

$$L(Y, f(X)) = E[(Y - f(X))^2] \tag{1}$$

In this paper, we use the square root of Eq. (1), the *root mean squared error* (RMSE). The criterion is used to estimate the parameters of the underlying model (e.g., the parameters in a GLM may be estimated by the Newton-Rhapson method that minimizes Eq. (1)). The "trained" model is the model with parameters chosen to minimize RMSE, which is estimated by computing the empirical loss function.

Once the model is trained, it should be validated to estimate how good it is at making predictions. The goal is to estimate how the model will perform (in terms of prediction error) on new data examples. A model may perform very well on the training data while performing very poorly on new data, a problem called overfitting [7]. For instance, we train a GLM on FEMA 156 data so that we can predict retrofit cost for a building that we are interested in retrofitting (in other words, a building that is not in the training data). Validation is the key step to assessing generalizability of a model. Performance is measured by estimating RMSE on data that is *not used to train the model*. Otherwise, RMSE estimates will be biased because the model already knows the data! When the training data is not very large, as in our case, cross-validation can provide unbiased estimates of RMSE [8].

In the following subsections, we describe the particular machine learning methods we use in this paper, including the least complex method: the linear regression model.

2.1 (Generalized) Linear Model

In classical linear regression, the outcome is typically assumed to follow a normal distribution. The generalized linear model (GLM) extends the classical linear regression model by allowing the outcome to follow any distribution in the exponential family, such as the normal, binomial, Poisson, and gamma distributions [10]. Depending on the outcome distribution, the GLM can be used for regression or classification problems (for instance, the binomial distribution yields logistic regression, a widely used statistical model for binary classification).

As the name suggests, the GLM is essentially linear. While the GLM is not strictly linear, it nevertheless imposes a restrictive structure on the relationship between features and outcome. One might wonder if this restriction somehow hurts performance, in terms of high prediction error. On the other hand, as discussed in Fung et al. [1], the GLM has several strengths: relative to a more complex model such as a neural network, it is both easier to train and easier to interpret. The latter is particularly important if the ultimate objective is not just to predict Y but to understand what features are driving the prediction.

2.2 Ensemble Methods: Bagging and Boosting Trees

Tree-based methods, such as Classification and Regression Trees (CART) are a very different type of model to the GLM. At a high level, a decision tree partitions the feature space into different segments and uses the mean value of the outcome corresponding to a chosen segment. For instance, if X represents building age, we partition X into two segments, say $X \leq 50$ and $X > 50$. Given X_{new}, we predict Y_{new} based on the mean value of Y in the segment that X_{new} belongs to. If X represents multiple features, say building age and building historic status, then partitioning is recursive as shown in Fig. 1 (hence the tree structure) [11].

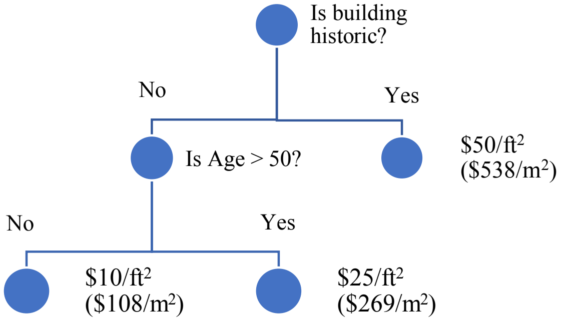


Fig. 1 – Example of a regression tree with two features: historic status of building and age of building. The terminal nodes represent predictions based on the mean retrofit cost in that branch of the tree. For instance, for a non-historic building that is over 50 years old, the predicted cost is $25 per square foot.

Representation of recursive partitioning as a tree makes these models interpretable, since they reflect the steps of a decision-making process. Moreover, the structure has the potential to easily capture interactions between features as well as non-linear relationships between features and outcome. Unfortunately, the hierarchical structure comes with a significant disadvantage: trees often have high variance [8]. Intuitively, this is because trees are highly sensitive to the partitioning rule, since any errors in the partitioning will be propagated down the tree. As a result, prediction error from a single tree can be large.

One way around these limitations is to combine the predictions of many trees. Ensemble methods leverage multiple individual algorithms to obtain better performance than each algorithm on its own. Two of the most common ensemble methods are *bagging* and *boosting*.

Bagging, short for bootstrap aggregating, resamples the training data (with replacement) B times, fitting the model to each of the B models and averaging the predictions across the B models. This process reduces the variance from fitting a single model to the training data without increasing the bias [12]. While bagging was originally applied to trees, the method can be applied to create an ensemble of any model. Bagging trees is robust to noise, outliers, and avoids overfitting. In this paper, we use *random forests*, a bagging method that additionally uses random subsamples of the features in order to reduce correlation across trees [13].

Boosting, on the other hand, fits a model sequentially M times to the training data. At each step m, the training data is reweighted based on the model's performance in the previous step, with the goal of focusing learning on the training examples that the model is predicting poorly, and at the last step M the final prediction is a weighted average of the M predictions [14]. Boosting was originally proposed for classification but has since been extended to regression. In contrast to bagging, boosted trees are not robust to noise or outliers. In this paper, we use a *gradient boosted model* (GBM), a boosting method that addresses some of the weaknesses of boosting, namely speed and robustness [15].

While boosting superficially resembles bagging, its objective makes the ensemble procedure very different. Boosting iteratively improves the performance of a "weak" learner (i.e., learner with high prediction error) to create a "strong" learner that has lower prediction error than the weak learners. Bagging, on the other hand, is a special case of model averaging: each model is fit independently of the others.

As is immediately obvious, the performance improvements provided by bagging and boosting come at a cost: combining multiple trees means the final "model" loses its interpretability. The question we ask in this paper is whether the improvement in performance (ie, lower prediction error) justifies the loss of interpretability: in other words, not just if such models are better but rather how much better.

2.3 Neural Networks

Another set of models that has recently received a lot of attention is the neural network. Neural networks are not new, however. The foundational theory dates back to the 1940s, with a mathematical model of how the brain processes complex information through the use of many simple neurons (hence the name) [18]. What is relatively novel is that the combination of large amounts of data and computing power in the early 21st century has allowed such previously intractable algorithms to be used broadly.

Despite the name and original motivation, neural networks have not been successful for plausibly modeling biological systems. Instead, neural networks have found success in statistical pattern recognition. The quintessential neural network is the feedforward neural network, or the multi-layer perceptron (MLP), essentially a "function approximation machine" [19]. A feedforward network learns a mapping $Y = f(X)$ by passing the inputs X through multiple hidden layers of neurons. Each layer finds a new representation of the features that is fed to the next layer, as shown in Fig. 1. Thus, the function f that is learned is simply a composition of many simpler functions. The term "feedforward" refers to information going in one direction (from inputs to output). If the output is fed back into the model, the network is called *recurrent*. Feedforward networks are most suitable for structured, tabular data [18].

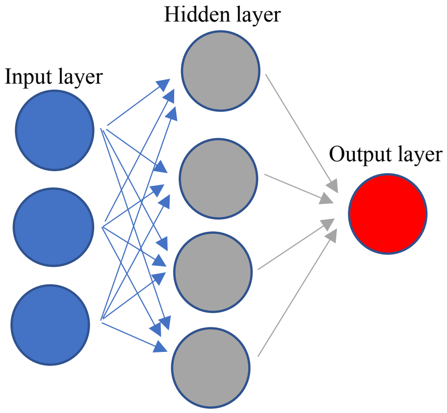


Fig. 2 – Example of a feedforward neural network with a single hidden layer. The input layer represents the features, which get passed on to the hidden layer. The hidden layer transforms the input features into a set of hidden features, which are passed on to the output layer to produce predictions. Note that the number of neurons in the hidden layer does not have to be the same as the number of features.

This simple idea belies the complexity of feedforward neural networks. The number of layers, including input layer, hidden layers, and output layer, defines the depth of the network. This is where the term deep learning originates. Moreover, the hidden layers can perform nonlinear transformations of the inputs and thus the network can learn a nonlinear and complex function f that is difficult to learn with a regression model.

Indeed, a neural network may be thought of as a nonlinear generalization of linear regression [8]; if the units in the hidden layers all perform an "identity" transformation, the entire model collapses to a linear regression model.

Estimation of the underlying parameters, however, requires solving a nonlinear, non-convex optimization problem. Fortunately, the gradients can be computed efficiently through *backpropagation* [18]. A thorough treatment of neural networks is far beyond the scope of this paper. For a deeper dive, see Goodfellow [19].

Like the ensemble methods discussed in Sec. 2.2, the complexity of a feedforward neural network makes interpretability (understanding how the algorithm produced a prediction) virtually impossible. Indeed, the impressive performance of neural networks has been criticized for coming at the expense of "black box learning." The interpretability problem has become a hot research topic in artificial intelligence (AI) more generally, as evidenced by the nascent field of Explainable AI; see Adadi and Berrada [20] for a survey.

3. The Training Data

In this section, we describe the training data used to predict seismic retrofit costs, originally collected for FEMA 156 [9]. The data is freely available online as part of FEMA's archived Seismic Rehabilitation Cost Estimator (SRCE) software (SRCE [16]). The data set includes retrofit cost and building characteristics for each of 1978 buildings in the United States and Canada. While the data set does not include primary building use, it nevertheless represents the most complete publicly available data set on seismic retrofit costs for North American buildings. The SRCE data is discussed in detail in Fung et al. [1].

In this paper, we use the features presented in Table 1. As shown in the Table, we focus on predicting *structural* retrofit cost (in particular, we predict the structural *unit* cost: structural cost per square foot). As we show in Fung et al. [1], predicting total retrofit cost is associated with much higher prediction error.

Table 1 – Description of the SRCE training data used in this paper

Variable	Definitions	Range of values
Structural retrofit cost	Cost to retrofit structural elements of a building, in millions of US dollars	[0.002, 157.161]
Area	Building floor area, in 1000ft$_2$ (93m$_2$)	[0.2, 1430.3]
Age	Building age, in years since construction	[2, 153]
Stories	Number of above and below ground stories	[1, 38]
Historic	Is building deemed historic?	Y, N
Occupancy	What happens to occupants during retrofit?	IP, TR, V
Seismicity	Site seismicity	L, M, H, VH
Performance objective	Retrofit target performance objective	LS, DC, IO
Building type	Building model type	See Table 2

Costs in the original SRCE data are normalized to average construction costs in California for 1993. Following Fung et al. [1], we normalize costs to US national average construction costs in 2016, using the Engineering News Record's Building Construction Index (BCI) [17]. It is worth noting that retrofit engineering practice has evolved since the SRCE data was collected, likely decreasing the rate of growth in retrofit costs relative to the growth in the material and labor costs represented in the BCI.

Historic is a binary variable denoting whether the building is deemed historic and, thus, requires special care to preserve historic elements of the building. Occupancy denotes whether occupants are left in-place (IP), temporarily relocated to another part of the building (TR), or completely vacated (V) during the retrofit. Seismicity is measured using peak ground acceleration (*pga*) with a 5 % chance of exceedance in 50 years and is grouped into four categories [1]. In principle, any measure that captures the variation in seismic exposure will work. The performance objectives in the SRCE data are Life Safety (LS), the most common target, Damage Control (DC), and Immediate Occupancy (IO). For definitions, see FEMA 156 [9].

Finally, Table 2 presents the 15 building types represented in the SRCE data. In the models, we group building types based on structural similarities [1]. Note that, by far, unreinforced masonry is the most highly represented building type (32 %), followed by concrete frame with infill walls (C3) and concrete shear wall (C2), each at about 18 %. More details on the SRCE data are given in Fung et al. [1].

Table 2 – Building groups, building types, and their fraction in the SRCE data

Group	Building type	Building type name	Fraction in data
1	URM	Unreinforced Masonry	31.85 %
2	W1	Wood Light Frame	3.28 %
	W2	Wood (Commercial or Industrial)	4.85 %
3	PC1	Precast Concrete Tilt Up Walls	3.34 %
	RM1	Reinforced Masonry with Metal or Wood Diaphragm	5.24 %
4	C1	Concrete Moment Frame	7.54 %
	C3	Concrete Frame with Infill Walls	18.22 %
5	S1	Steel Moment Frame	4.98 %
6	S2	Steel Braced Frame	2.29 %
	S3	Steel Light Frame	1.31 %
7	S5	Steel Frame with Infill Walls	7.86 %
8	C2	Concrete Shear Wall	17.96 %
	PC2	Precast Concrete Frame with Infill Walls	0.98 %
	RM2	Reinforced Masonry with Precast Concrete Diaphragm	0.85 %
	S4	Steel Frame with Concrete Walls	2.11 %

4. Results

This section presents the results of training five different models on the SRCE data. In particular, we consider a linear regression model, a generalized linear model (GLM), a bagging model (random forests), a boosting model (GBM), and a neural network model (MLP).

We train each of the five models on the SRCE data using K-fold cross-validation in order to estimate out-of-sample performance, with $K = 10$. K-fold cross-validation randomly splits the data into K mutually exclusive subsets, iteratively using each as a validation set while training the model on the remaining data. Out-of-sample prediction error is estimated by averaging RMSE across the K iterations. Hyperparameters (e.g., tree depth, number of hidden units) are chosen using random grid search. Table 3 presents out-of-sample prediction error estimates, as well as training error (that is, average RMSE from training a model), for each of the models. The results suggest that the neural network, GBM, and random forest outperform the linear regression models. In particular, the prediction error for the linear regression model is \$3/ft$_2$ (\$33/m$_2$) higher than for the neural network.

Table 3 – Out-of-sample prediction error estimates (validation error), with their standard deviation. Based on K-fold cross-validation with $K = 10$. The training error represents the minimization of the loss function (in this case, RMSE), used to estimate the model parameters. Values in 2016 USD per ft$_2$ (m$_2$).

Model	Validation error	Standard deviation	Training error
Neural network	34.99 (376.64)	9.16 (98.60)	28.66 (308.50)
GBM	35.07 (377.50)	9.67 (104.09)	13.48 (145.39)
Random forest	35.14 (378.26)	9.93 (106.89)	36.27 (390.42)
GLM	37.13 (399.68)	9.10 (97.95)	38.57 (415.18)
Linear regression	38.04 (409.47)	8.59 (92.47)	38.63 (415.82)

Does the \$3/ft$_2$ (\$33/m$_2$) reduction in prediction error justify the loss in interpretability? Ultimately, this is subjective and depends on the problem context and the extent of the retrofit in terms of square footage. For a building owner that only cares about obtaining the most accurate estimate, the tradeoff may be highly valuable. In particular, for a building owner facing a 10,000ft$_2$ (929m$_2$) project, the potential cost from larger error is \$30,000. In contrast, for a consultant or any other user that must be able to provide an explanation as to *why* the cost estimate is y, a GLM may be desirable as it provides transparency and interpretability as to what features are driving the cost estimate. In this case, a \$2/ft$_2$ (\$21/m$_2$) reduction in prediction error may not be so appealing, especially for a smaller retrofit project.

4.1 Model Details

For completeness, we now describe each of the model architectures. The GLM uses a gamma distribution for the outcome with a log link function, i.e., $\ln(E[Y|X]) = X\beta$ [1]. The neural network includes one hidden layer with 500 neurons, with a rectified linear unit (ReLU) activation function. (The activation function transforms the weighted sum of the input to the layer into an output for the next layer; the ReLU activation function is the standard activation function for MLPs as it typically achieves better performance than other functions due to its handling of the vanishing gradient problem. For more details, see Goodfellow et al. [19].) The number of hidden layers and neurons is chosen by random grid search.

The GBM and random forest models can best be described in terms of number of trees and tree depth (i.e., the number of splits an individual tree performs), which are hyperparameters chosen by random grid search. For the GBM, the number of trees (that is, the number of boosting iterations) is M=46, with a maximum depth of 14. This results in an average of 547 leaves (i.e., terminal or decision nodes), with the smallest tree having only 174 leaves. The random forest model fits 37 separate trees, with a maximum depth of 29. This results in an average of 690 leaves, with the smallest tree having 550 leaves.

Each of the five models employs a form of *regularization*, a technique used to avoid overfitting when training a model. For the linear regression model and the GLM, regularization takes the form of adding a penalty term to the objective function. Broadly, regularization penalizes model complexity. The linear regression model in Table 3 uses a quadratic penalty (this is commonly called ridge regression) [8]. The GLM, on the other hand, employs a linear penalty term (known as lasso). For the neural network, we use *dropout* for regularization [21]: in the hidden layer, 50% of the neurons are randomly dropped before passing the inputs to the activation function.

Regularization for ensembles of trees, on the other hand, is to a certain extent "built in." For GBM, the most straightforward method is to limit the number of boosting iterations M. For random forests, limiting tree depth is the easiest method of regularization. It is worth noting that in the random forest model trained in Table 3, all 37 trees have depth 29. For more advanced regularization techniques, see Hastie et al. [8].

It is worth noting that the neural network in Table 3 is not very deep, as it only contains a single hidden layer. A deeper neural network, with three hidden layers (each with 200 neurons, ReLU activation functions, and 40% dropout) performs worse than the neural network in Table 3, with out-of-sample prediction error of 37.28. While it does outperform the GLM and linear regression model, these results demonstrate that a more complex model is not always better.

5. Conclusion

In this paper, we considered the relative performance of machine learning methods, including linear regression models, ensembles of decision trees, and neural networks, as well as the tradeoff between prediction error and interpretability, in the context of predicting seismic retrofit costs. The results suggest that the gain in performance from more complex machine learning methods such as neural networks may not be significant to justify the loss of interpretability when a cost estimate may require some level of transparency. If the objective is to obtain an accurate estimate, especially when the building is very large, a neural network with a single hidden layer can reduce prediction error by \$3/ft$_2$ (\$33/m$_2$). Perhaps encouragingly, these results suggest that nonlinear relationships between features and the outcome are not hugely important for predicting retrofit cost.

It is worth noting that the ensemble methods we use combine the predictions of multiple "weak" learners in order to create a "strong" learner with better performance. An alternative ensemble method called *stacking* combines "strong" learners to create a "super learner" [22, 23]. An example of a super learner would be to stack random forests, GBMs, and neural networks into a single learner. Of course, the super learner would be inscrutable in terms of interpretability since each of the individual learners lacks interpretability. Nevertheless, it would be interesting to compare such a super learner to the models we consider in this paper. This is left for future work.

6. Acknowledgements

We are grateful to Shane Crawford and Stanley Gilbert for comments. All errors are our own.

7. Copyrights

8. References

[1] Fung JF, Butry DT, Sattar S, McCabe SL (2020). A predictive modeling approach to estimating seismic retrofit costs. *Earthquake Spectra*, **36** (2) *Forthcoming*. DOI: 10.1177/8755293019891716.

[2] Jafarzadeh R, Wilkinson S, Gonzalez V, Ingham J, Amiri GG (2013a). Predicting seismic retrofit construction cost for buildings with framed structures using multilinear regression analysis. *Journal of Construction Engineering and Management*, **140** (3) 04013062.

[3] Jafarzadeh R, Ingham J, Wilkinson S, Gonzalez V, Aghakouchak A (2013b). Application of artificial neural network methodology for predicting seismic retrofit construction costs. *Journal of Construction Engineering and Management*, **140** (2) 04013044.

[4] Nasrazadani H, Mahsuli M, Talebiyan H, Kashani H (2017). *Probabilistic modeling framework for prediction of seismic retrofit cost of buildings*. *Journal of Construction Engineering and Management*, **143** (8) 04017055.

[5] Elfaki AO, Alatawi S, Abushandi E (2014). Using intelligent techniques in construction project cost estimation: 10-year survey. *Advances in Civil Engineering*, **2014** 1-11.

[6] Mitchell T (1997): *Machine Learning*. McGraw Hill, 1st edition.

[7] Murphy KP (2012): *Machine Learning: A Probabilistic Perspective*. The MIT Press, 1st edition.

[8] Hastie T, Tibshirani R, Friedman J (2009): *The Elements of Statistical Learning*. Springer, 2nd edition.

[9] FEMA (1994): Typical costs for seismic rehabilitation of existing buildings, Volume 1: Summary. *FEMA 156*, Federal Emergency Management Agency, Washington, DC, USA.

[10] Nedler JA, Wedderburn RWM (1972): Generalized Linear Models. *Journal of the Royal Statistical Society. Series A*, **135** (3) 370-384.

[11] Breiman L, Friedman J, Olshen R, Stone C (1984): *Classification and Regression Trees*. CRC Press.

[12] Breiman L (1996): Bagging predictors. *Machine Learning*, **24** (2) 123-140.

[13] Breiman L (2001): Random forests. *Machine Learning*, **45** (1) 5-32.

[14] Schapire RE (2003): The boosting approach to machine learning: An overview. *Nonlinear Estimation and Classification*. Springer.

[15] Friedman JH (2001): Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, **29** (5) 1189-1232.

[16] FEMA (2013-2014): SRCE: Seismic Rehabilitation Cost Estimator. https://www.fema.gov/media-library/assets/documents/30220. Accessed: 2016-10-15.

[17] ENR (2017): Engineering News Record: Historical Indices. https://www.enr.com/economics/historical_indices. Accessed: 2017-03-03.

[18] Bishop CM (2006): *Pattern recognition and machine learning*. Springer. https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/.

[19] Goodfellow I, Bengio Y, Courville A (2016): *Deep Learning*. MIT Press. http://deeplearningbook.org.

[20] Adadi A, Berrada M (2018): Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, **6** 52138-52160. DOI: 10.1109/ACCESS.2018.2870052.

[21] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014): Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15** 1929-1958. http://jmlr.org/papers/v15/srivastava14a.html.

[22] Breiman L (1996): Stacked regressions. *Machine Learning*, **24** (1) 49-64.

[23] van der Laan MJ, Polley EC, Hubbard AE (2007): Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1) 1544-6115.

Evaluation of an alternative null detector for adapted Wheatstone bridge

Shamith U. Payagala*, Alana M. Dee†, and Dean G. Jarrett*

*National Institute of Standards and Technology, 100 Bureau Drive, Stop 8171, Gaithersburg, MD, 20899, USA
shamith.payagala@nist.gov

†University of Pittsburgh, Swanson School of Engineering, Pittsburgh, PA, 15261, USA

***Abstract* — The adapted Wheatstone bridge technique has been utilized at National Institute of Standards and Technology (NIST) and other National Metrology Institutes (NMIs) for high resistance measurements. In this work, we evaluate the suitability of a transimpedance amplifier (TIA) as a null detector for the adapted Wheatstone bridge to improve high resistance measurements. Initial resistance measurements using the TIA based null detector showed an order of magnitude improvement in the Type A uncertainty.**

***Index Terms* — High resistance measurement, low-current measurements, resistance standards, transimpedance amplifier.**

I. Introduction

At the National Institute of Standards and Technology (NIST) automated adapted Wheatstone bridges [1] are used for accurate high resistance measurements from 10 MΩ to 100 TΩ. An adapted Wheatstone bridge uses two voltage sources in place of the main resistance ratio arms of a conventional Wheatstone bridge. High value resistors (≥ 1 MΩ) are used in the other ratio arms. At the balance point, the resistance value of the unknown resistor can be calculated using the equation:

$$R_x = R_s(V_1/V_2) \tag{1}$$

Multiple variants of the adapted Wheatstone bridge have been developed at NIST [2]. To achieve a bridge balance, the adapted Wheatstone bridge uses a commercial electrometer as the current detector. Adapted Wheatstone bridges at NIST utilize current null detection [3]. The third generation of the adapted Wheatstone bridge implemented at NIST balances at a true null rather than $\pm\ 5000 \times 10^{-6}$ from null as the first and second generation of these bridges at NIST [4]. Balancing at a true null rather than at a null offset, combined with well calibrated voltage sources, provides scaling independent of Hamon transfer standards that use the substitution technique. In addition, these improvements to the bridge have allowed the extension of the measurement range to higher decades of resistance (> 10 TΩ) than Hamon transfer standards can achieve [5]. Balancing the bridge at a true null requires low current measurements in the pA range. The accuracy and the measurement noise of the commercial electrometer has been identified as a possible area of improvement to reduce Type A uncertainties of high resistance measurements. Hence, alternative methods of low current detection were examined.

II. Evaluation of a transimpedance amplifier

A transimpedance amplifier (TIA) converts low level current to measurable voltage. Thus, a TIA, combined with a precision multimeter, provides the means for accurate low current measurements. To examine the suitability of a TIA as a current null detector, a commercially available variable gain sub femto-Ampere current amplifier was evaluated. To use the TIA in the adapted Wheatstone bridge, a recently calibrated 8-1/2 digit multimeter was used as the voltage readout. The TIA of interest (Femto DDPCA300)[1] has 3 adjustable low pass filter settings ranging from full bandwidth, 0.7 Hz, and 0.1 Hz. The gain can be adjusted over a wide range from 10^4 up to 10^{13} V/A [6].

A. Linearity

Measurements were made to study the linearity of the TIA as a current detector. Prior current measurements made during the bridge balance have identified ±2 pA as the operational current range of the third generation adapted Wheatstone bridge with true null balancing. Using a source-meter as an input current source, the output voltage of the TIA was measured for a set of currents to within ±2 pA. The voltage output of the TIA was used to calculate the output current based on the transimpedance, which is then compared with the input current of the source-meter. For the specific current range of interest, gain settings above 10^{10} V/A showed a maximum deviation of 10 fA from the linear regression. Thus, gain settings above 10^{10} V/A were considered for successive tests.

B. Optimal Gain and Bandwidth

The 10^4 to 10^{13} V/A gain range of the TIA provides the means for high resolution current measurements for a wide measurement range (1 mA to sub-fA). However, considering the specific use case of the adapted Wheatstone bridge, the gain must be optimized for the specific current range of ±2 pA that is commonly measured during the bridge balance.

Using a source-meter as an input current source, the voltage output of the TIA was measured using an 8-1/2 digit multimeter. Sets of currents within ±2 pA was sourced for gain

[1] Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Payagala, Shamith; Dee, Alana; Jarrett, Dean G. "Evaluation of an alternative null detector for adapted Wheatstone bridge." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 22, 2020 - August 28, 2020.

settings 10^{10} V/A, 10^{11} V/A, and 10^{12} V/A respectively. For each gain setting, 160 voltage measurements were made using 100 power line cycles and the last 120 measurements are accounted for statistics. Based on the standard deviation of each set of voltage measurements, the 10^{10} V/A gain setting produced the lowest measurement noise from the three gain settings tested.

Similarly, the three bandwidth settings of the TIA were evaluated individually based on the measurement noise of a known input current. Using the same setup as the gain setting tests, the voltage output of the multimeter was recorded for a set of currents within ±2 pA. Using the 10^{10} V/A gain setting based on previous results, measurements were repeated for the three different bandwidth settings. Measurement results suggest that the lowest bandwidth setting of 0.1 Hz is the optimum case for DC low current measurements of the adapted Wheatstone bridge.

III. Improvements to the Bridge Measurements

The goal of using the TIA as a current detector is to reduce the systematic uncertainties of high resistance measurements. The third generation of the adapted Wheatstone bridge at NIST was modified to use the TIA, along with the 8-1/2 digit multimeter, as a null detection method. To study improvements of using the TIA as a null detector, a resistance ratio of 10:1 using well known 100 GΩ and 10 GΩ resistors were employed. A gain of 10^{10} V/A and a bandwidth of 0.1 Hz for the TIA was used based on the prior measurements.

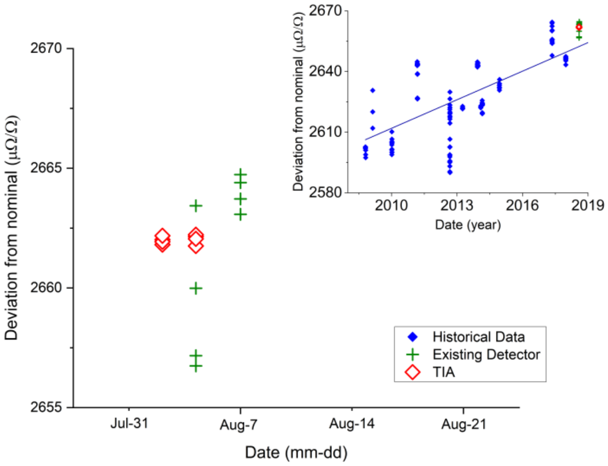


Fig. 1. 100 GΩ resistance measurement comparison between TIA and existing current detector. Historical data with a linear regression fit is shown in the inset.

To compare the repeatability and the accuracy of the TIA as a null detector, multiple resistance measurements were made using both TIA and the existing detector. Fig. 1 shows a set of data acquired using both detectors. Historical data for the 100 GΩ resistor that dates to 2008 are shown in the inset. A linear regression line based on the historical data was used to predict the value of the resistor at the time of the measurement. Data points in green (symbol +) using the existing detector has a standard deviation of 3.2 µΩ/Ω whereas the data points in red (symbol ◇) acquired by the TIA resulted in a standard deviation of 0.2 µΩ/Ω. The average of both sets of data has an agreement of 1 µΩ/Ω, implying both detectors produced comparable average resistance value while the TIA resulted in improved repeatability. Reported expanded uncertainty for 100 GΩ standard resistor is 50 µΩ/Ω [7].

During a resistance measurement, the balancing algorithm reverses the polarity of the voltage sources to reduce bridge offsets. Examining the currents resulted at bridge balance for both polarities, it was evident that the TIA produced significantly lower current offsets compared to the existing detector. The settling times were comparable for both detectors.

IV. Conclusion

A commercially available TIA was evaluated as a current null detector for the adapted Wheatstone bridge. Measurements made using a known current input yielded satisfactory linearity between the input and the TIA output. The optimum gain setting of 10^{10} V/A and the bandwidth of the TIA for the specific current range of interest ±2 pA was realized. Preliminary resistance measurement comparison between the TIA and the existing current detector suggests an order of magnitude improvement in the Type A uncertainty.

References

[1] L. C. A. Henderson, “A new technique for the automated measurement of high valued resistors,” *J. Phys. Electron. Sci. Instrum*., vol. 20, pp. 492 – 495, 1987.

[2] D. G. Jarrett, “Automated guarded bridge for calibration of multi-megohm standard resistors from 10 MΩ to 1 TΩ,” *IEEE Trans. Instrum. Meas*., vol. 46, no. 2, pp. 325-328, April 1997.

[3] G. Rietveld and J.H.N. van der Beek, “Automated High-Ohmic Resistance Bridge with Voltage and Current Null Detection”, *IEEE Trans. Instrum. Meas*., vol. 62, no. 6, pp. 1760-1765, 2013.

[4] D. G. Jarrett, S. U. Payagala, M. E. Kraft, K. M. Yu, “Third Generation of Adapted Wheatstone Bridge for High Resistance Measurements at NIST,” CPEM 2016 Conf. Digest, July 2016.

[5] D. G. Jarrett, A. J. Dupree, "Next generation guarded Hamon transfer standards for high resistance," CPEM 2010 Conference Digest, June 13-18, 2010, Daejeon, South Korea, pp. 571-572.

[6] FEMTO Messtechnik GmbH, “Variable Gain Sub Femto Ampere Current Amplifier,” DDPCA-300 datasheet, May 2016.

[7] R. Elmquist, D. Jarrett, G. Jones, M. Kraft, S. Shields and R. Dziuba. “NIST Measurement Service for DC Standard Resistors”, NIST Technical Note 1458, December 2003.

Over-The-Air Calibration of a Dual-Beam Dual-Polarized 28-GHz Phased-Array Channel Sounder

Sung Yun Jun, Derek Caudill, Jelena Senic
RF Technology Division
National Institute of Standards and Technology
Boulder, CO, USA
{sungyun.jun; derek.caudill; jelena.senic}@nist.gov

Camillo Gentile, Jack Chuang, Nada Golmie
Wireless Networks Division
National Institute of Standards and Technology
Gaithersburg, MD, USA
{camillo.gentile; jack.chuang; nada.golmie}@nist.gov

***Abstract*—This paper describes an over-the-air (OTA) calibration procedure for the impulse response of a 28-GHz phased-array antenna channel sounder. The silicon-germanium (SiGe) antenna board is composed from two 8x8 planar arrays; each array can generate a distinct beam in the vertical and horizontal polarizations that can be steered in both azimuth and elevation within ±50° and ±25°, respectively. As part of the calibration procedure, unique pre-distortion filters were designed per steer angle, achieving a peak-to-sidelobe ratio in the impulse response of up to 56.5 dB, with less than 1 dB variation across the steer angles. Critically, we show that performance degraded up to 10.1 dB when applying the pre-distortion filter designed at one steer angle to the other steer angles, underscoring the need for angle-specific filters. Finally, we show that performance between both polarizations is comparable thanks to the array symmetries in the fabrication process.**

Keywords—channel sounder; 5G wireless communications; propagation channel; calibration; wireless communication

I. Introduction

The millimeter-wave (mmWave) frequency bands are drawing attention amid growing interest for fifth-generation (5G) communications [1]. While these operating frequencies enable high-bandwidth and high-speed links, they suffer from greater propagation loss. To overcome the latter, phased-array antennas paired with beam-steering to achieve high-gain directional systems have been recently considered for 5G wireless networks [2].

Previously, channel sounders at the mmWave frequency bands of 28 GHz, 60 GHz, and 83 GHz based on time-multiplexed switched antenna arrays were developed by the National Institute of Standards and Technology (NIST) [3]-[4]. The calibration procedure to estimate and remove systematic distortions caused by the IF (Intermediate frequency) and RF (Radio frequency) sections of the system was based on a back-to-back method in which waveguide attenuators were directly connected between the transmitter (TX) and receiver (RX) after removing the antennas. When dealing with phased-array antennas, the back-to-back method cannot be applied because the RF components in the SiGe chips on the printed circuit board – which also distort the system impulse response – are not accessible through connectors. As such, OTA methods must be used instead.

In this paper, we extend the work in [5] to a 3D-steerable dual-polarized 28-GHz phased-array antenna. In Section II, the system architecture of the 28-GHz channel sounder is described following by the OTA calibration procedure and results in Section III. Lastly, the conclusion and the future work are described in Section IV.

II. Channel Sounder Architecture

Our 28-GHz dual-beam dual-polarized phased-array channel sounder is a correlation-based system using a pseudo-random noise (PN) code as the probing signal. Fig. 1 shows the schematic block diagram of the channel sounder calibration. An arbitrary waveform generator (AWG) at TX provides the binary phase shift keying modulated IF signal of 2.5 GHz using a PN code with degree of 11 (2047 chips). A chip rate of 1 Gbits/s with 2 GHz null-to-null bandwidth is used. The IF frequency is up-converted to RF with center frequency at 28.5 GHz. The received RF signal is down-converted back to IF of 2.5 GHz and digitized using an analog-to-digital converter (ADC) sampling at 10 Gsample/s. A single 10 MHz rubidium-disciplined crystal oscillator is used as the frequency reference. Two rubidium devices are used when channel sounding to enable untethered time synchronization of PN codes between TX and RX.

Details of the 28 GHz dual-beam dual-polarized phased-array in our system are described in detail in [6]. The array consists of 2x64 patch elements driven by sixteen 2x4 SiGe dual-beamformer transmit/receive (TRX) chips, placed on a low-cost printed-circuit board (PCB) and operated with 6 bits of phase control and 22 dB of gain control. The scan angle range of the array with vertical- (V) and horizontal- (H) polarized beams in the azimuth and elevation planes is ±50° and ±25°, respectively. Given the limited azimuthal scan range of each board, four boards – each scanning ±45° – are mounted at right angles with respect to each other in order to extend the total scan to an omni-directional view (360°). The array can generate V- and H-polarized beams simultaneously.

III. Impulse-response calibration

In order to obtain optimal performance, we calibrate the system impulse response through pre-distortion filtering to account for the hardware non-idealities of the channel sounder. As seen in previous work [4]-[5], the pre-distortion filter is designed in a multi-step procedure: In the first step, the ideal PN code is transmitted and the code distorted by the system is

Jun, Sung Yun; Caudill, Derek; Senic, Jelena; Gentile, Camillo; Chuang, Jack; Golmie, Nada T. "Over-The-Air Calibration of a Dual-Beam Dual-Polarized 28-GHz Phased-Array Channel Sounder." Paper presented at 2020 IEEE International Symposium on Antennas and Propagation and North American Radio Science Meeting, Montreal, CA. July 05, 2020 - July 10, 2020.

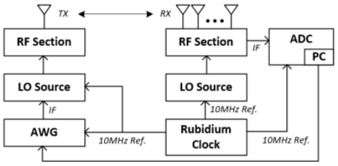


Fig. 1. Block diagram of the channel sounder calibration setup.

received; next, the received code is deconvolved from the ideal code, what we refer to as the *pre-distorted code*; finally, the pre-distorted code is sent instead of the ideal code so that what is received after distortion by the system is a quasi-ideal code. The calibration procedure was applied iteratively to deal with the non-linearities of the system; three iterations were necessary for convergence to optimal performance. Details of the calibration procedure can be found in [4].

Fig. 2 displays the OTA calibration set-up: It employs a scalar feed horn at the TX with 16.6 dBi gain and a phased-array board at the RX. The RX was placed in a semi-anechoic environment on a precision mechanical rotator with 1.6 m separation from the TX; the purpose of the rotator was to align the antenna boresights while investigating different phased-array steer angles. Fig. 3(a) displays the normalized power delay profile (PDP) with and without pre-distortion filtering for the V-polarized beam at 0° azimuth steer angle in the elevation plane and Fig. 3(b) for H-polarized beam; thanks to array symmetries in the fabrication process, the results for both polarizations are comparable. Calibration greatly reduced the peak Interval Of Discrimination (IOD_{pk}) between the correlation peak and the sidelobes: in the correlation tail, non-ideal spurious peaks were down at least 54.8 dB in both polarizations; also, the cross-polarization rejection between beams was measured as approximately 22.7 dB and 24.4 dB in V- and H-polarized beams, respectively.

Table I compares the IOD_{pk} for the response calibrated at several azimuth steer angles (in the elevation plane) versus the uncalibrated response: the IOD_{pk} is similar across the steer angles. In contrast, the pre-distortion filter designed at 0° azimuth steer angle was then applied to the other steer angles, what we refer to as a *semi-calibrated response*: the IOD_{pk} suffered by up 10.1 dB, suggesting that calibration per steer angle is essential. Cross-polarization rejection remained within 1 dB of original values.

IV. Conclusion and Future work

This paper presents results for the impulse-response calibration of a dual-beam dual-polarized 28-GHz phased-array channel sounder. The study shows that the results, expressed in terms of the peak-to-sidelobe correlation ratio, are essentially invariant across the various steer angles of the phased-array antenna so long as a unique calibration is performed per steer angle; in contrast, we show that when a single calibration is performed and merely applied to other steer angles, the ratio can suffer by up to 10.1 dB. Finally, we show that results are similar between the V- and H-polarized beams due to the array symmetries in the fabrication process. So far, we have only investigated performance with variation in the azimuth plane; in future work, we shall investigate joint variation in the azimuth and elevation planes.

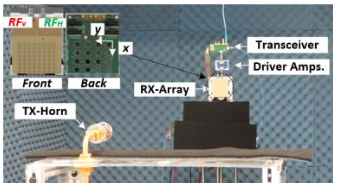
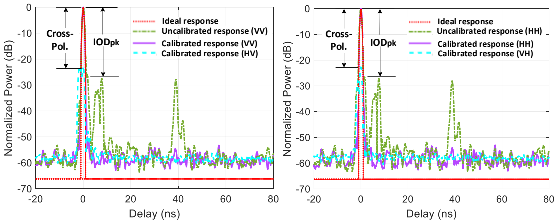


Fig. 2. Measurement setup for the OTA impulse response calibration.



(a) (b)

Fig. 3. Normalized PDP with uncalibrated and calibrated response at 28.5 GHz at boresight (a) V-polarized beam (b) H-polarized beam.

TABLE I. IOD_{PK} (IN DB) THE ELEVATION PLANE

Steer Angle (Azimuth)	*Uncalibrated Response*		*Calibrated Response*		*Semi-Calibrated Response**	
	VV	**HH**	**VV**	**HH**	**VV**	**HH**
-22.5°	27.2	26.9	55.6	54.9	48.1	48.1
-13.5°	27.5	26.9	55.3	54.9	46.2	51.9
-4.5°	27.3	27.0	55.2	55.0	47.3	52.0
0°	27.2	27.0	54.8	54.1	54.8	54.1
4.5°	27.4	27.1	55.5	55.7	48.3	53.3
13.5°	27.3	27.0	56.5	53.4	46.4	47.9
22.5°	27.3	27.1	55.2	54.4	49.9	53.5

* Pre-distortion filter designed at 0° azimuth steer angle and applied to other steer angles.

References

[1] T. Rappaport, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!", *IEEE Access*, vol. 1, pp. 335-49, 2013.

[2] C. Gentile, P. B. Papazian, N. Golmie, K. A. Remley, P. Vouras, J. Senic, J. Wang, D. Caudill, C. Lai, R. Sun, and J. Chuang, "MillimeterWave Channel Measurement and Modeling: A NIST Perspective," *IEEE Communications Magazine*, vol. 56, no. 12, pp. 30–37, Dec 2018.

[3] P. B. Papazian, C. Gentile, K. A. Remley, J. Senic, N. Golmie, "A radio channel sounder for mobile millimeter-wave communications: System implementation and measurement assessment", *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 9, pp. 2924-2932, Sep. 2016.

[4] R. Sun, P. B. Papazian, J. Senic, Y. Lo, J.-K. Choi, K. A. Remley, and C. Gentile, "Design and calibration of a double-directional 60 GHz channel sounder for multipath component tracking," in *Proc. IEEE European Conf. Antennas and Propagation*, pp. 1–5, Mar. 2017.

[5] D. Caudill, P. B. Papazian, C. Gentile, J. Chuang and N. Golmie, "Omnidirectional Channel Sounder With Phased-ArrayAntennas for 5G Mobile Communications," *IEEE Trans. Microw. Theory Techn.*, vol. 67, no. 7, pp. 2936-2945, July 2019.

[6] A. Nafe, M. Sayginer, K. Kibaroglu and G. M. Rebeiz, "2x64 Dual-Polarized Dual-Beam Single-Aperture 28 GHz Phased Array with High Cross-Polarization Rejection for 5G Polarization MIMO," in *2019 IEEE MTT-S International Microwave Symposium (IMS), Boston*, MA, USA, 2019, pp. 484-487.

Evaluation of NMIJ Traveling Dual Source Bridge Using NIST Adapted Wheatstone Bridge

T. Oe[1, 2], S. Payagala[2], D. G. Jarrett[2], and N.-H. Kaneko[1]
[1]National Metrology Institute of Japan (NMIJ), Advanced Industrial Science and Technology (AIST), Tsukuba, 305-8563, Japan
t.oe@aist.go.jp
[2]National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, 20899, USA

***Abstract*—DC high resistance measurement capability between National Institute of Standards and Technology (NIST) and National Metrology Institute of Japan (NMIJ) has been evaluated using a NMIJ traveling dual source bridge. The NMIJ bridge determines the resistance ratio by measuring the voltage ratio using an 8.5 digit digital multimeter. Based on the same 1 MΩ standard resistor calibrated using a 2-terminal cryogenic current comparator, standard resistors from 10 MΩ to 10 GΩ were calibrated by 10:1 scaling measurements with both systems. The measurements at 10 GΩ agreed to less than 1 µΩ/Ω for both systems. Scaling from 10 GΩ to 100 TΩ is in progress.**

***Index Terms*—adapted Wheatstone bridge, dual source bridge, DC resistance standard, high resistance standard, small current measurement.**

I. Introduction

National Institute of Standards and Technology (NIST) and National Metrology Institute of Japan (NMIJ) have performed high resistance measurement comparisons for several years, which yielded good agreement within the expanded uncertainties ($k = 2$) for 10 TΩ and 100 TΩ [1]. For the high resistance measurement comparisons, the NMIJ traveling dual source bridge (DSB) was sent to NIST where high resistance measurements were performed with the NMIJ DSB and the third generation of the NIST adapted Wheatstone bridge (AWB). Both measurement systems are based on the method of the automated high resistance bridge proposed by Henderson [2] but use different hardware and calibration methods. The NMIJ traveling DSB relies on the linearity of an 8.5 digit digital multimeter [3], which measures the voltage ratio applied to standard resistors to calculate the resistance ratio. The NIST AWB #3 system uses well characterized voltage sources to apply a voltage corrections to derive the correct resistance ratio. The high resistance comparison was performed starting with a stable 1 MΩ standard resistor calibrated using the 2-terminal cryogenic current comparator bridge at NIST [4], [5]. This recently calibrated 1 MΩ resistor was used in both high resistance measurement systems as the initial reference. The comparison built up to 10 GΩ by repeating 10 : 1 ratio measurements, the calibrated 10 GΩ values of both systems agreed to within 1 µΩ/Ω. Scaling from 10 GΩ to 100 TΩ will be completed in the next several months. In this abstract, both measurement systems are introduced and the comparison results up to 10 GΩ are shown briefly.

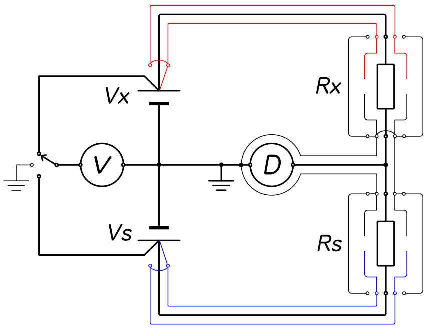


Fig. 1. Circuit diagram of the NMIJ traveling dual source bridge.

II. Measurement system

A. NMIJ Traveling High Resistance Bridge

Figure 1 shows the circuit diagram of the NMIJ bridge. It uses two 6.5 digit voltage sources (ADCMT 6166)[1], a current detector, a digital multimeter (Hewlett Packard 3458A) to measure the voltage ratio, and a switch box using latching relays. The latching relays were switched using a USB I/O device with LabVIEW software. All equipment is controlled by the laptop PC connected through a USB isolator to prevent noise problems. The flowing current between the neutral points of the bridge was converted to voltage using the transimpedanse amplifier (Femto DDPCA-300) and the output voltage was measured using a 7.5 digit digital multimeter (Keithley DMM7510). The low terminal of the DMM7510 was connected to the same ground as the chassis of all equipment that were connected via power cables grounded to the same node. An active guarding was implemented by applying V_x and V_s to the cable shields and the split guard of the resistance boxes as shown in Fig. 1 to shorten the settling time and to make the measurements more stable.

The bridge is balanced by adjusting V_s so that the current detector achieves null current. Then the voltage ratio was measured by measuring V_x and V_s with the 3458A. The detailed algorithm is shown below;

[1] Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Oe, Takehiko; Payagala, Shamith; Jarrett, Dean G.; Kaneko, Nobu-Hisa. "Evaluation of NMIJ Traveling Dual Source Bridge Using NIST Adapted Wheatstone Bridge." Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

1) Apply nominal ratio voltages (*e.g.* $V_\mathrm{x}/V_\mathrm{s} = +100\ \mathrm{V}/-10\ \mathrm{V}$) and the voltage V_s is adjusted to determine the bridge balance voltage ($V_{\mathrm{s}0+}$).
2) Repeat the same process for reversed polarity ($-100\ \mathrm{V}/+10\ \mathrm{V}$) and determine $V_{\mathrm{s}0-}$.
3) Adjust V_s based on $V_\mathrm{x}/V_\mathrm{s} = +100\ \mathrm{V}/(V_{\mathrm{s}0+}+V_\mathrm{off})$, then the flowing current is measured using the DDPCA-300 and the DMM7510. Next, the voltage ratio is measured using the 3458A. Here, the V_off is indicated voltage offset and the typical value is from 0.01 mV (1 µV/V of 10 V) to 1 mV (100 µV/V of 10 V).
4) 3) is repeated for $V_\mathrm{x}/V_\mathrm{s} = -100\ \mathrm{V}/(V_{\mathrm{s}0-}+V_\mathrm{off})$, $+100\ \mathrm{V}/(V_{\mathrm{s}0+}-V_\mathrm{off})$, and $-100\ \mathrm{V}/(V_{\mathrm{s}0-}-V_\mathrm{off})$.
5) 3) and 4) are repeated for indicated number of loop.

The output of the voltage source V_x was adjusted so that the absolute values of the voltage measured by the 3458A was exactly 100 V. We use the assumption that the absolute value of ±100 V is the same for both polarities. To measure the voltages V_x and V_s by the 3458A, the same voltage range was used for the 3458A to achieve an accurate voltage ratio. The 3458A input is connected to the ground during the current measurement and it is connected to V_x or V_s only during the voltage measurement. The voltage ratio measurement was performed before and after the current measurement to cancel the effect of the drift of the actual applied voltage.

B. NIST Adapted Wheatstone Bridge (AWB) #3

The NIST AWB#3 uses two FLUKE 57X0A calibrators as the stable voltage sources, and a Keithley 6430A as a current detector [6]-[8]. The output voltages of the 57X0As are calibrated annualy using a Measurement International 8000A/8001A automated Cutkosky binary voltage divider. The voltage corrections are applied at the time of measurement to V_x and V_s. Two versions of the control software were available (LabVIEW and Visual Basic) which showed similar results. The Visual Basic algorithm was used for the comparison since the settling time was the same as that of the AIST system. Detailed measurement algorithm is shown below:

1) The applied voltage to get the bridge balance ($V_{\mathrm{s}0+}$ and $V_{\mathrm{s}0-}$) is calculated from the predicted value of the R_s and R_x using the historical data.
2) $V_\mathrm{x}/V_\mathrm{s} = +100\ \mathrm{V}/(V_{\mathrm{s}0+}+V_\mathrm{off})$ is applied and the flowing current is measured using the 6430A. Here, the V_off is 5000 µV/V typically.
3) 2) is repeated for $V_\mathrm{x}/V_\mathrm{s} = -100\ \mathrm{V}/(V_{\mathrm{s}0-}+V_\mathrm{off})$, $+100\ \mathrm{V}/(V_{\mathrm{s}0+}-V_\mathrm{off})$, and $-100\ \mathrm{V}/(V_{\mathrm{s}0-}-V_\mathrm{off})$.
4) 2) and 3) are repeated for indicated number of loop.

The NMIJ traveling DSB adopts similar algorithm as the NIST AWB#3. The difference of the systems is that the NIST AWB#3 uses the predicted value to determine the V_s to get the bridge balance ($V_{\mathrm{s}0+}$ and $V_{\mathrm{s}0-}$), and the NMIJ DSB determines the values from the brief measurement at the beginning of the measurement.

III. Measurement results

Table I summarises the measured results by the NMIJ traveling DSB and the NIST AWB#3. The two middle columns show the measured relative resistance values, and the right column shows the difference of the two systems in µΩ/Ω. The same 1 MΩ value were used as the reference standard for the 10 MΩ measurement for both systems. For the higher resistance measurements, each system used its own calibrated value for the reference resistor (R_s) for each step, so the difference values show the accumulated error. The maximum difference of the resistance ratio was -0.3 µΩ/Ω in the 10 GΩ/1 GΩ measurement. The NMIJ's calibration and measurement capability (CMC) value for 10 GΩ in the key comparison data base (KCDB) is 6.3 µΩ/Ω and a difference of -0.6 µΩ/Ω is 10 % of it, based on this, the results were satisfactory.

TABLE I
Measured relative resistance value from 10 MΩ to 10 GΩ by the NIST AWB#3 and the NMIJ traveling DSB.

Nominal resistance value	NMIJ traveling DSB [µΩ/Ω]	NIST AWB #3 [µΩ/Ω]	Difference (NMIJ - NIST) [µΩ/Ω]
10 MΩ	17.4	17.5	-0.1
100 MΩ	23.2	23.3	-0.1
1 GΩ	-580.8	-580.5	-0.3
10 GΩ	4249.0	4249.6	-0.6

IV. Conclusion

We have performed the high resistance comparison from 10 MΩ to 10 GΩ by repeating 10:1 measurements. The calibrated 10 GΩ value determined from scaling with each system agreed within 1 µΩ/Ω, and it is sufficiently low to use the dual source bridge as a traveling standard. In the future, a comparison is going to be performed up to 100 TΩ and the NMIJ's voltage injection type Wheatstone system in Japan for the customer calibration [9] will be evaluated using the traveling DSB.

References

[1] D. G. Jarrett, T. Oe, N.-H. Kaneko, and S. U. Payagala, ”10 TΩ and 100 TΩ resistance comparison between NIST and AIST,” *CPEM 2018 Conf. Digest*, July 2018.

[2] L. Henderson, “A new technique for the automatic measurement of high value resistors,” *J. Phys. E: Sci. Instrum.*, vol. 20, pp. 492-495, September 1987.

[3] J. I. Giem, ”Sub-ppm linearity testing of a dmm using a Josephson junction array,” *IEEE Trans. on Instrum. and Meas.*, vol. 40,no. 2, pp. 329–332, April 1991.

[4] F. L. Hernandez-Marquez, M. E. Bierzychudek, G. R. Jones Jr., and R. E. Elmquist, “Precision high-value resistance scaling with a two-terminal cryogenic current comparator,” *Rev. Sci. Instrum.*, vol. 85, 044701, April 2014.

[5] R. E. Elmquist, E. Hourdakis, D. G. Jarrett, and N. M. Zimmerman, “Direct resistance comparisons from the QHR to 100 MΩ using a cryogenic current comparator,” *IEEE Trans. Instrum. Meas.*, vol. 54, pp. 525-528, April 2005.

[6] D. G. Jarrett, S. U. Payagala, M. E. Kraft, K. M. Yu, “Third generation of adapted Wheatstone bridge for high resistance measurements at NIST,” *CPEM 2016 Conf. Digest*, July 2016.

[7] D. G. Jarrett, “Evaluation of guarded high resistance Hamon transfer standards,” *IEEE Trans. Instrum. Meas.*, vol. 48, no. 2, pp. 324-328, April 1999.

[8] D. G. Jarrett and M. E. Kraft, “10 TΩ and 100 TΩ resistance measurements at NIST,” *X-Semetro Congress. Digest*, Buenos Aires, Argentina, Sept. 25-27, 2013.

[9] T. Oe, J. Kinoshita, and N.-H. Kaneko, “Voltage injection type high ohm resistance bridge,” *CPEM 2012 Conf. Digest*, pp. 360-361, June 2012.

Oe, Takehiko; Payagala, Shamith; Jarrett, Dean G.; Kaneko, Nobu-Hisa. ”Evaluation of NMIJ Traveling Dual Source Bridge Using NIST Adapted Wheatstone Bridge.” Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

Direct DC Voltages Comparison between two Programmable Josephson Voltage Standards at SCL

S. Yang*, S. Cular†, A. Rüfenacht†, C. J. Burroughs†, P. D. Dresselhaus†, S. P. Benz† and M. N. Ng*

*Standards and Calibration Laboratory
Hong Kong, China
steven.yang@itc.gov.hk

†National Institute of Standards and Technology
Gaithersburg, MD 20899 and Boulder, CO 80305, USA

***Abstract* — Standards and Calibration Laboratory (SCL) setup in 2018 a liquid helium based programmable Josephson voltage standard (PJVS) developed by the National Institute of Standards and Technology (NIST). The system was validated by direct comparison with a NIST transportable PJVS system at 1.018 V, 4 V, 6 V, 8 V and 10 V. The difference between the systems was within 0.5 nV, with an expanded uncertainty of less than 2.2 nV (*k* = 2). In this paper, the setup and the results of the direct comparison method are presented.**

***Index Terms* — Programmable Josephson voltage standard, quantum voltage measurement, measurement uncertainty, Josephson voltage comparison.**

I. Introduction

In 1993, the Standards and Calibration Laboratory established the 1 V conventional Josephson array voltage standard (CJVS) as the primary standard of DC voltage for Hong Kong. Then the system was upgraded to 10 V in 1995. This conventional system was designed based on the National Institute of Standards and Technology (NIST) design. The system operates through a computer control interface "NISTVOLT" developed by NIST. The microwave signal was generated by a Gunn-diode oscillator at a frequency around 75 GHz. The SCL maintains 10 V Josephson array chips developed by both NIST and Physikalisch-Technische Bundesanstalt (PTB). The array chip was cooled by liquid helium stored in a 100 liter dewar.

The recent evolution of quantum voltage standard technology, namely the programmable Josephson voltage standard (PJVS), enables the generation of both dc voltage and stepwise approximated ac voltages for frequency up to around 1 kHz [1 – 2]. The PJVS system developed by NIST can produce intrinsically stable DC voltages up to 10 V at a microwave driving frequency of about 18-20 GHz.

In 2018, the SCL setup a new PJVS to upgrade the 25-year-old CJVS system. The performance of this SCL PJVS system was validated by direct comparison with a NIST transportable PJVS system, as shown in Fig. 1, and by indirect comparison with the in-service SCL CJVS system through the use of a set of Zener voltage standards.

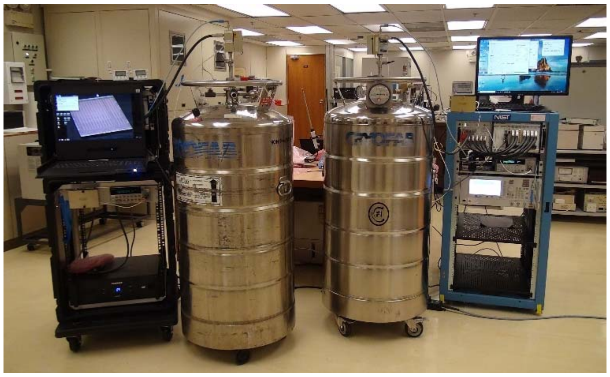


Fig. 1. Direct comparison between NIST transportable PJVS (left) and SCL PJVS (right)

II. PJVS Comparison Setup

A. SCL PJVS system

The SCL PJVS system is a liquid-helium-based system developed by NIST. The PJVS array has 23 subarrays ranging in size from 6 junctions up to 16 800 junctions. The total number of junctions with all subarrays in series is 248 312, such that the entire circuit is capable of generating 10 V with a driving frequency of approximately 20 GHz. The PJVS array bias electronics is a 24-channel current source that employ 16-bit DACs to supplies bias currents to the 23 subarrays. A microwave signal generator and a microwave power amplifier were installed to supply driving frequencies in the range of (18 – 20) GHz with sufficient power to drive all the subarrays. A digital nano-voltmeter was used as a null detector on its 1 mV range. The measurement range of the null detector was fixed to minimize noise and to avoid the change in gain when different ranges are used. The system operation is controlled by a NIST program.

B. NIST Transportable PJVS System

The NIST transportable PJVS system is similar in fundamental design and operation to the SCL PJVS system; however, the system has been implemented with a compact RF

Yang, S.; Cular, Stefan; Rufenacht, Alain; Burroughs, Charles J.; Dresselhaus, Paul; Benz, Samuel P.; Ng, M. "Direct DC Voltages Comparison between two Programmable Josephson Voltage Standards at SCL." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

source and installed in rack cases for easy palletization and shipment. The system additionally has a frequency-to-voltage converter that enables easy operation from power supplied at any common line voltage and frequency without reconfiguration.

C. Measurement configuration at SCL

The measurement was conducted on site at the direct current laboratory of the SCL at Wan Chai, Hong Kong, China from March 12 to March 16, 2018. During the measurement, the frequency sources of both PJVS systems were locked to the same 10 MHz signal supplied by the RF laboratory of the SCL, which is traceable to the laboratory's cesium-beam frequency standards and is compared routinely with the BIPM through GPS common view comparison.

Various grounding conditions were tested prior to the start of comparisons. Both systems were connected to the same power outlet and grounded at a single point. Both systems were then verified to be stable during the course of measurement through the use of hourly quantum locking range [3] checks.

III. Measurement Results

A. Direct comparison results

Automated measurement was performed by the NIST program in the following order: 10 V, 1.018 V, 4 V, 6 V, 8 V and 0 V. At least 200 polarity pairs of measurement data were collected at each test voltage. The measured difference between the two PJVS systems and their expanded uncertainty at different test voltages are tabulated at Table I. Overall, the difference between the two systems was within 0.5 nV, with an expanded uncertainty of within 2.2 nV ($k = 2$).

Table I. Measured voltage differences

Test Voltage	Mean difference between PJVS systems	Expanded Uncertainty ($k = 2$)
10 V	0.2 nV	2.2 nV
8 V	0.2 nV	2.0 nV
6 V	0.4 nV	1.9 nV
4 V	0.3 nV	1.8 nV
1.018 V	-0.5 nV	1.8 nV

The uncertainty budget was intended to be conservative for this comparison and was composed of eight parameters with the three largest contributors being the digital null meter noise (0.8 nV), frequency sources (0.7 nV) and the null meter non-linearity (0.6 nV). In some direct JVS comparisons, through the use of a common 10 MHz signal and similar RF synthesizers, the frequency uncertainty is removed from the budget. However, for this comparison the sources were of different design and it was chosen to include the values.

B. Leakage resistance to ground measurement

For improved precision in direct PJVS comparison, NIST custom-made polytetrafluoroethylene (PTFE) isolated cables were used between the DACs and the amplifier board, and between the cryoprobe and the amplifier board [4]. The leakage current between two precision leads to ground were measured at 10 V using the same setup for the NIST–BIPM 10 V PJVS comparison [4 - 5]. The resulting leakage current to ground of the SCL PJVS system at 10 V is tabulated in Table II. The overall leakage resistance to ground can be calculated by the sum of the measured leakage current (144 pA), as it can be assumed that the system is virtually biased at the same voltage of 10 V. The corresponding leakage resistance to ground of the SCL PJVS system was 69 GΩ, which is much better than the required leakage resistance specification of 25 GΩ.

Table II. Measured leakage Current to ground at 10 V

Configuration	Leakage current to ground
(+) side to ground	105 pA
(-) side to ground	39 pA
Overall	144 pA

IV. Conclusion

The on-site comparison of two PJVS systems was conducted at SCL for five voltages. The overall result was a difference of 0.2 nV with an expanded uncertainty of 2.2 nV ($k = 2$) at 10 V. This comparison with a transportable PJVS system demonstrated nano-volt level uncertainties. Uncertainties were decreased through the use of low-leakage custom PTFE insulated cables.

References

[1] C. J. Burroughs, P. D. Dresselhaus, A. Rüfenacht, D. Olaya, M. M. Elsbury, Y. Tang and S. P. Benz, "NIST 10 V Programmable Josephson Voltage Standard System," *IEEE Trans. Instrum. Meas.,* vol. 60, no. 7, pp. 2482 – 2488, July 2011.

[2] Y. Tang, V. N. Ojha, S. Schlamminger, A. Rüfenacht, C. J. Burroughs, P. D. Dresselhaus and S. P. Benz, "A 10 V programmable Josephson voltage standard and its applications for voltage metrology," *Metrologia,* vol. 49, pp. 635 – 643, September 2012.

[3] A. Rüfenacht, N. E. Flower-Jacobs and S. P. Benz, "Impact of the latest generation of Josephson voltage standards in ac and dc electric metrology," *Metrologia,* vol. 55, S152, 2018.

[4] A. Rüfenacht, C. J. Burroughs, P. D. Dresselhaus and S. P. Benz, "Measurement of Leakage Current to Ground in Programmable Josephson Voltage Standards," *CPEM 2018*, July 2018.

[5] S. Solve, A. Rüfenacht, C. J. Burroughs and S. P. Benz, "Direct comparison of two NIST PJVS systems at 10 V," *Metrologia,* vol. 50, pp. 441 – 451, August 2013.

Yang, S.; Cular, Stefan; Rufenacht, Alain; Burroughs, Charles J.; Dresselhaus, Paul; Benz, Samuel P.; Ng, M. "Direct DC Voltages Comparison between two Programmable Josephson Voltage Standards at SCL." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

PAPER ID: 1137
DOI: 10.18462/iir.rankine.2020.1137

Working Fluid Screening for Ocean Thermal Energy Conversion (OTEC) Applications

Ian BELL[(a)]

[(a)]National Institute of Standards and Technology
Boulder, CO, 80305, USA, ian.bell@nist.gov

ABSTRACT

Ocean Thermal Energy Conversion (OTEC) is among the small number of renewable energy production systems that can provide base load capacity. Ammonia has long been proposed and tested as a working fluid for these systems. *Are there any other superior fluids from a purely thermodynamic standpoint?* This question was probed based on the use of highly accurate thermodynamic equations of state, and a few other possible working fluids have been identified, but many candidates have disqualifying features: acute toxicity, high global warming potential, high ozone depletion potential, extreme flammability or very low condensing pressures. No identified fluids were obviously superior to ammonia, though CO_2 and water deserve further study.

Keywords: Ocean Thermal Energy, Rankine Cycle, Renewable Energy.

1. INTRODUCTION

While there are significant technical and cost challenges with OTEC systems that have limited their application, a comprehensive study of working fluids for OTEC applications has not been carried out thus far. Existing studies (Yoon, et al., 2014; Sun, et al., 2012; Moore & Martin, 2008) have usually studied only a few working fluids at a time, and here the goal is to test a broader palette of working fluids, relaxing some of the current constraints on suitability of working fluids for this application. A rather comprehensive review of OTEC systems is available on Wikipedia.

One major challenge with OTEC is the curse of the second law -- the small temperature difference between the surface water and the subsurface water limits the maximum efficiency possible to that of a Carnot cycle given by $\eta_{\rm Carnot} = 1 - T_l/T_h$, where T_l is the low temperature reservoir and T_h the high temperature reservoir. Large thermocline gradients allow for temperature differences greater than 20 K in the equatorial Pacific Ocean between the sea surface and the water at a depth of 1 km, but generally, the temperature differences are much smaller. A typical assumption (surface seawater: 25 °C, subsurface seawater: 5°C), yields a Carnot efficiency of 6.7%; all thermodynamically feasible systems will operate at a fraction of this efficiency.

2. ANALYSIS

2.1. Cycle Modeling

The component modelling employed in this study is intentionally simplified, with the focus placed on the thermodynamics of the OTEC cycle. The OTEC system is embodied by a conventional closed four component Rankine cycle, in which subsurface seawater cools the condenser and surface water heats the evaporator. Figure 1 shows the overall schematic. The modeling assumptions are that: a) the pressure drop in the heat exchangers are negligible (reasonable as plate or shell-and-tube heat exchangers are likely to be used) b) fixed outlet pinch from boiler of 1 K c) fixed outlet pinch of 1 K in condenser d) fixed overall isentropic efficiency in pump and turbine (including motor and generator losses, respectively) of 0.8 e) inlet seawater to boiler at 25 °C, inlet seawater to condenser at 5 °C.

The only adjustable parameters in the cycle optimization are the superheat at the outlet of the boiler (which controls the evaporation saturation temperature), and the subcooling at the outlet of the condenser (which controls the condensation saturation temperature). These two temperature differences are allowed to float in

Bell, Ian. "Working Fluid Screening for Ocean Thermal Energy Conversion (OTEC) Applications." Paper presented at IIR International Rankine 2020 Conference - Heating, Cooling and Power Generation, Glasgow, UK, Glasgow, UK. July 26, 2020 - July 29, 2020.

the range 0.1 K to 10 K, and Figure 2 makes clear that these temperature differences should be as small as possible to minimize heat transfer irreversibilities, as expected. Therefore the two temperature difference terms were set to 0.1 K.

In the seawater loop, the pressure drops in the seawater pipes were assumed to be zero. This assumption is certainly invalid, but the assumption is employed for all working fluids, so as to make that part of the analysis a fair comparison between fluids.

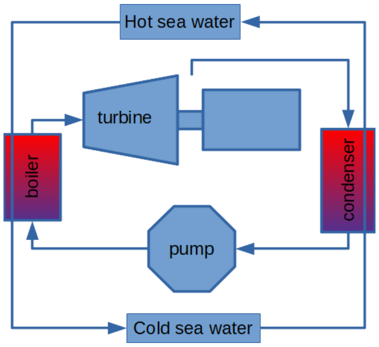


Figure 1 Schematic of OTEC system

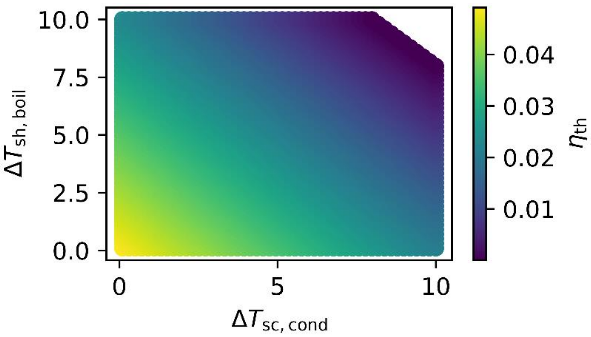


Figure 2 Thermal efficiency of ammonia OTEC system as function of temperature differences in HX

2.2. Thermodynamic Modeling and Constraints

The thermodynamic properties were obtained from the NIST REFPROP library, version 10.0 (Lemmon, et al., 2018) and REFPROP was called by CoolProp (Bell, et al., 2014); the cycle model was written in Python. REFPROP implements the thermodynamic models that are deemed to be the most reliable in the literature. Due to space constraints the references for each equation of state cannot be included; please see the documentation for REFPROP for more information. In the first screening, all fluids were included, even some fluids that are clearly unsuitable for OTEC application (e.g., very toxic species like hydrogen sulfide). There are approximately 150 fluids included in REFPROP. Recent studies (McLinden, et al., 2017) have shown that there are few chemical compounds that would make suitable working fluids for air conditioning applications, and similar constraints apply here too.

One constraint that is sometimes invoked in working fluid selection for Rankine cycles is that the condenser saturation pressure should be above atmospheric pressure. This constraint eliminates associating fluids like water and alcohols, but these associating fluids are excellent working fluids from a thermodynamic standpoint, so that constraint was relaxed here. Furthermore, fluids with large global warming potential (GWP) or non-zero ozone depletion potential (ODP) might be legislatively constrained in the present or the future, but they were also allowed through the initial screening. Fluids known to be toxic were also included.

2.3. Screening results

Given the simplicity of the cycle modelling, only a few figures of merit can be considered, among which are the thermal efficiency and the volumetric output power. As should perhaps not be surprising, and in agreement with existing studies, one of the best working fluids for this application is ammonia. Therefore, the results for all other fluids are considered in relation to ammonia. The thermal efficiency quantifies how much output power can be obtained for a given input thermal input heat from the hot seawater, and also quantifies how close to the Carnot limit the cycle is able to approach.

In this very simplified cycle model, the ammonia thermal efficiency is 4.9%, or 0.7 times the Carnot thermal efficiency. The inclusion of additional loss parameters (e.g., pressure drop in the seawater pipes), and increased irreversibilities in heat transfer, mechanical systems, and electro-mechanical conversion would have the effect of further reducing the thermal efficiency of the system, highlighting the challenge of designing OTEC systems that have competitive economics with technologies like photovoltaics or wind power, even though those technologies are fundamentally intermittent while OTEC has the promise of base load generation.

The volumetric output power (VOP), defined as the mass-specific output power times a density gives some indication of the relative size of components that would be required. For instance, in the context of cooling systems, it indicates how large the compressor would need to be (Bell, et al., 2019), and similar arguments apply here. A very large volumetric output power indicates that the pipes could be designed with a reasonable pressure drop for a given output capacity and also gives some rough indication of how large the heat exchangers would need to be to ensure a feasible pressure drop in them. The quantity VOP is tightly linked with the operating pressure of the system -- higher-pressure working fluids will result in higher densities at all points in the cycle, and therefore, higher volumetric output power. In this study, the volumetric output power was determined based on the density of the working fluid at the inlet to the turbine.

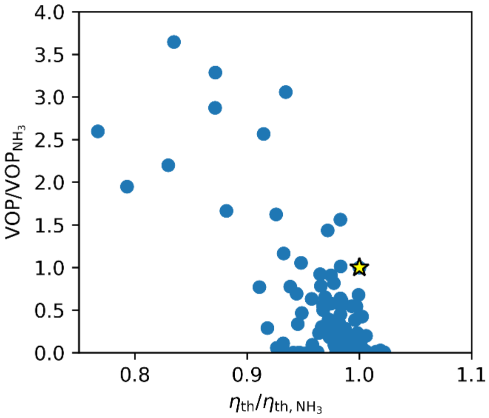


Figure 3 Scatter plot of normalized VOP versus normalized η_{th} for candidate working fluids in REFPROP 10.0. The star indicates ammonia (NH_3)

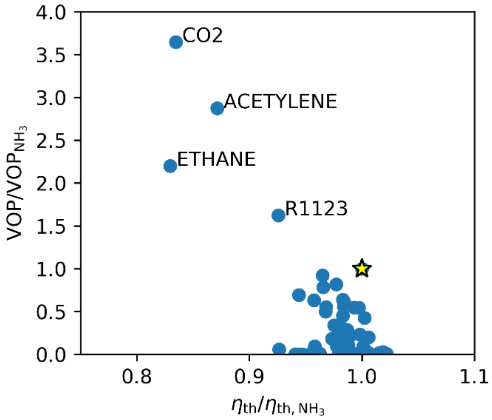


Figure 4 Scatter plot as in Figure 3, but screened for toxicity, GWP, and ODP

Figure 3 shows a scatter plot of the normalized volumetric output power and normalized thermal efficiency for OTEC systems, with each metric normalized by the value for ammonia. In this screening there are no working fluids that have simultaneously a larger volumetric output power and higher thermal efficiency than ammonia, but if either of the figures of merit are relaxed, some new candidate working fluids emerge. Fluids which are acutely toxic (e.g, chlorine, vinyl chloride, ethylene oxide) would not be suitable working fluids because of the danger to any humans or animals that might become exposed to the fluid during manufacture or service. A similar argument applies to fluids with high ODP. The screening was now limited to fluids without known acute toxicity, having an ODP less than 10, and having a global warming potential less than 100. If the GWP or ODP was not known, the fluid was retained. The updated version of Figure 3 with these screening criteria applied is presented in Figure 4.

Of the fluids with higher VOP than ammonia, the only one that seems like a promising candidate is CO_2. Two candidates are extremely flammable (acetylene and ethane), and the other (R-1123) doesn't offer a markedly improved VOP than ammonia. Carbon dioxide has not been discussed as a candidate working fluid for OTEC applications, perhaps because its critical temperature is rather low (31 °C), but that has not precluded the recent interest in supercritical CO_2 power cycles. Indeed, while the working pressures for CO_2 would be approximately 8 times higher than ammonia in the condenser (circa 40 bar versus circa 5 bar), this comes with an upside: high working pressures mean that the impact of pressure drop is less significant, CO_2 has excellent heat transfer properties, and the heat exchangers could be made rather compact, potentially improving the economics, which are generally dire for OTEC as compared with other renewable power generation technologies.

The VOP is intimately tied to the fluid through its working pressure, which are in turn tied to the range of seawater temperatures. The thermal efficiency shows a weaker dependence on the fluid -- the band of predicted

Bell, Ian. "Working Fluid Screening for Ocean Thermal Energy Conversion (OTEC) Applications." Paper presented at IIR International Rankine 2020 Conference - Heating, Cooling and Power Generation, Glasgow, UK, Glasgow, UK. July 26, 2020 - July 29, 2020.

VOP varies over a few orders of magnitude while the values of η_{th} do not vary nearly as widely. Thus once a decision has been made about the acceptable band of VOP, a more detailed study of component sizing and modelling can begin.

2.4. Discussion

Overall the predicted thermal efficiencies do not vary too widely among fluids, and the origin of the behaviour merits a brief comment. One of the assumptions invoked in the cycle modelling is that the pinching in the heat exchangers occurs at the outlet of the heat exchanger in both cases. As the seawater flow rates and temperature changes are not constrained in the cycle model, a simplifying assumption in the analysis here is that the seawater flow rates are infinite, which guarantees that the limiting capacitance is on the working fluid side, and therefore, the cycle can be schematically shown in temperature entropy coordinates (for ammonia in this case) as in Figure 5. This figure explains why the thermodynamic models for fluids operating well below their critical temperature consistently result in a similar thermal efficiency -- the dominant irreversibility in the cycle is that of heat transfer over a finite temperature difference in the boiler and the condenser, and if the condensing and boiling temperatures are fixed (as they are in this model), this loss will be roughly constant as a percentage of the output power for all fluids.

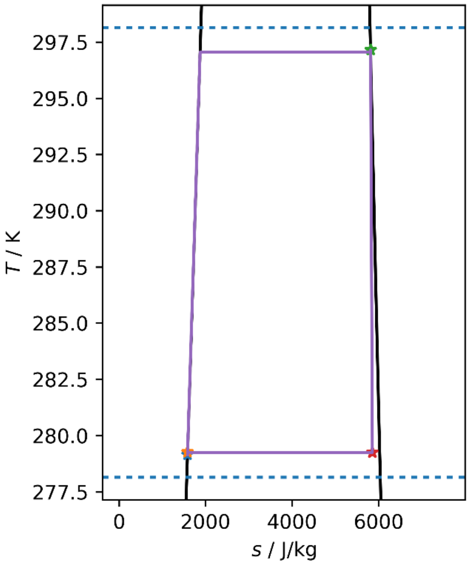


Figure 5 T-s plot for ammonia with cycle overlaid.

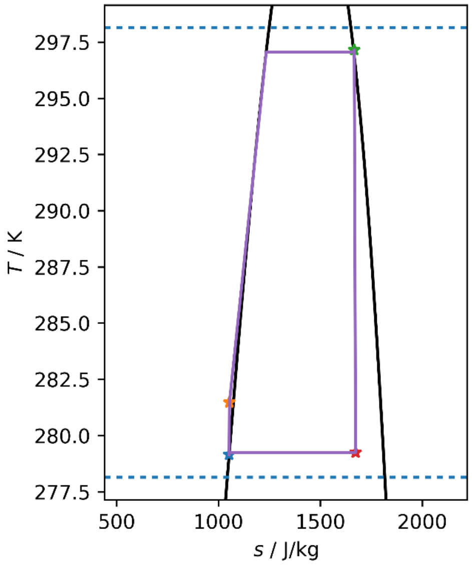


Figure 6 T-s plot for CO_2 with cycle overlaid.

The story is a bit different for CO_2; as shown in Figure 6, CO_2 boils at a pressure quite close to its critical pressure, and as a result, the thermal efficiency is lower because the cycle is more trapezoidal in these coordinates, deviating from the ideal rectangular Carnot cycle even more than fluids operating at lower pressures.

2.5. Summary

Table 1 summarizes the results for the fluids that survived the toxicity, GWP, and ODP screenings. A further screening was that the fluid should have a VOP that is at least one tenth of that of water. Many of the fluids are flammable (some are *very* flammable), and a predictive method could be used to estimate the flammability for the fluorinated species (Linteris, et al., 2019) if it were desired to include flammability in the screening.

Table 1 Fluids that survived the screening, sorted by VOP. Given names are the REFPROP identifier for each species.

Fluid	η_{th}	VOP / kJ m^{-3}
MDM	0.0470	0.49
C3CC6	0.0487	0.51
NONANE	0.0484	0.55
OXYLENE	0.0492	0.78
MXYLENE	0.0492	0.96
EBENZENE	0.0491	1.09
OCTANE	0.0484	1.59
D2O	0.0501	2.44
WATER	0.0502	2.74
TOLUENE	0.0493	2.93
MM	0.0473	4.34
C1CC6	0.0488	4.50
HEPTANE	0.0484	4.63
IOCTANE	0.0481	4.82
DMC	0.0491	5.59
ETHANOL	0.0498	6.67
R150	0.0494	7.50
BENZENE	0.0493	8.80
METHANOL	0.0500	12.95
HEXANE	0.0484	13.41
3METHYLPENTANE	0.0484	16.27
IHEXANE	0.0483	17.95
23DIMETHYLBUTANE	0.0483	19.34
ACETONE	0.0493	19.95
CYCLOPEN	0.0491	25.28
22DIMETHYLBUTANE	0.0481	25.32
NOVEC649	0.0455	26.64
PENTANE	0.0484	39.08
DEE	0.0484	42.13
RE347MCC	0.0470	43.82
R1336MZZZ	0.0480	44.59
1PENTENE	0.0485	47.14
IPENTANE	0.0483	50.1
R123	0.0485	52.0
R1233ZDE	0.0484	71.5
R1224YDZ	0.0481	81.1
NEOPENTN	0.0478	85.0
R1234ZEZ	0.0484	94.2
1BUTYNE	0.0486	98.9
C2BUTENE	0.0486	105.4
CYCLOBUTENE	0.0490	107.5
T2BUTENE	0.0484	113.5
BUTANE	0.0482	116.6
13BUTADIENE	0.0485	132.7
1BUTENE	0.0483	138.4
IBUTENE	0.0482	142.1
ISOBUTAN	0.0479	157.7
SO2	0.0492	197.6
CF3I	0.0483	209.2
R1234ZEE	0.0475	232.2
PROPYNE	0.0488	253.5
R1243ZF	0.0475	256.9
DME	0.0483	259.0
PROPADIENE	0.0483	284.4
R1234YF	0.0470	293.8
CYCLOPRO	0.0483	295.7
R1216	0.0464	322.5
PROPANE	0.0474	364.6
R161	0.0480	380.1
PROPYLEN	0.0474	429.0
AMMONIA	**0.0491**	**464.5**
R1123	0.0455	754.0
ETHANE	0.0407	1021.7
ACETYLENE	0.0428	1334.3
CO2	0.0410	1694.1

Bell, Ian. "Working Fluid Screening for Ocean Thermal Energy Conversion (OTEC) Applications." Paper presented at IIR International Rankine 2020 Conference - Heating, Cooling and Power Generation, Glasgow, UK, Glasgow, UK. July 26, 2020 - July 29, 2020.

3. CONCLUSIONS

The major conclusion of this study is that even after the most comprehensive fluid screening to date, there are no working fluids for ocean thermal energy conversion systems identified that are obviously superior to ammonia. Water and other associating fluids provide a slightly higher thermal efficiency, but their major downside is that the required components would be physically much larger than for ammonia. On the other hand, carbon dioxide appears to be an interesting fluid for this application due to its high working pressures, reasonable toxicity, and low GWP.

An important limitation of this study is that it involves solely thermodynamic optimization, not combined thermo-economic optimization (Quoilin, et al., 2011), which would almost certainly shift the optimal fluids from the screening. Most importantly, the heat exchanger analysis assumed a constant pinch temperature at the outlet of the working fluid. Achieving this fixed pinch would require widely varying heat exchanger configurations, and accordingly varying capital expenditure.

NOMENCLATURE

s	entropy (J kg^{-1})	T_h	temp., high (K)
T	temperature (K)	η_{th}	thermal efficiency
T_l	temp., low (K)	VOP	vol. output power (J m^{-3})

REFERENCES

Bell, I. H., Domanski, P. A., McLinden, M. O. & Linteris, G. T., 2019. The hunt for nonflammable refrigerant blends to replace R-134a. *Int. J. Refrig,* 8, Volume 104, pp. 484-495.

Bell, I. H., Wronski, J., Quoilin, S. & Lemort, V., 2014. Pure and Pseudo-pure Fluid Thermophysical Property Evaluation and the Open-Source Thermophysical Property Library CoolProp. *Ind. Eng. Chem. Res.,* Volume 53, pp. 2498-2508.

Lemmon, E. W., Bell, I. H., Huber, M. L. & McLinden, M. O., 2018. *NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 10.0, National Institute of Standards and Technology.*

Linteris, G. T., Bell, I. H. & McLinden, M. O., 2019. An empirical model for refrigerant flammability based on molecular structure and thermodynamics. *Int. J. Refrig,* Volume 104, pp. 144-150.

McLinden, M. O. et al., 2017. Limited options for low-global-warming-potential refrigerants. *Nat. Commun.,* Volume 8, p. 14476.

Moore, F. P. & Martin, L. L., 2008. A nonlinear nonconvex minimum total heat transfer area formulation for ocean thermal energy conversion (OTEC) systems. *Applied Thermal Engineering,* 6, Volume 28, pp. 1015-1021.

Quoilin, S., Declaye, S., Tchanche, B. F. & Lemort, V., 2011. Thermo-economic optimization of waste heat recovery Organic Rankine Cycles. *Applied Thermal Engineering,* 10, Volume 31, pp. 2885-2893.

Sun, F., Ikegami, Y., Jia, B. & Arima, H., 2012. Optimization design and exergy analysis of organic rankine cycle in ocean thermal energy conversion. *Applied Ocean Research,* 3, Volume 35, pp. 38-46.

Yoon, J.-I.et al., 2014. Efficiency comparison of subcritical OTEC power cycle using various working fluids. *Heat and Mass Transfer,* 2, Volume 50, pp. 985-996.

Ultrafast waveform metrology: A first international comparison

Mark Bieler[1], Paul Struszewski[1], Ari Feldman[2], Jeffrey Jargon[2], Paul Hale[2], Pengwei Gong[3], Wen Xie[3], Chuntao Yang[3], Zhigang Feng[4], Kejia Zhao[4], and Zhijun Yang[4]

[1] Physikalisch-Technische Bundesanstalt, Bundesallee 100, D-38116 Braunschweig, Germany, mark.bieler@ptb.de,
[2] National Institute of Standards and Technology, 325 Broadway, Boulder, CO 80305, USA, ari.feldman@nist.gov
[3] Beijing Institute of Radio Metrology and Measurement, 50 Yongding Road, Haidian District, Beijing, 100039, China, pwgong@qq.com
[4] National Institute of Metrology, 18 Beisanhuandonglu, Chaoyang District, 100029, Beijing, China, fengzg@nim.ac.cn

***Abstract* — We report on the first international comparison in ultrafast waveform metrology. To this end, the frequency and time responses of a photodiode with a nominal bandwidth of 100 GHz have been measured by several National Metrology Institutes during the last two years. An overall good agreement between the different measurement techniques is obtained. However, certain features in the frequency- and time-domain responses are not reproduced by all NMIs, calling for additional studies and comparisons.**

***Index Terms* — Waveform metrology, photodiodes, electro-optic sampling, ultrashort voltage pulses.**

I. Introduction

Ever increasing data rates in modern communication schemes pose the need for traceable measurements of electrical waveforms at mm-wave frequencies. Such *ultrafast waveform metrology* is currently developed at several National Metrology Institutes (NMIs) [1,2,3]. However, up to now, no detailed comparison of the different measurement procedures has been performed.

In this work, we report on the first international comparison in ultrafast waveform metrology. The device under test (DUT) is a photodiode (PD) with a nominal bandwidth of 100 GHz (u2t XPDV4120R)[1] with a coaxial 1.0-mm female connector and a 1-m long fiber for optical input at 1550 nm. NMIs from China (Beijing Institute of Radio Metrology and Measurement, BIRMM and National Institute of Metrology, NIM), USA (National Institute of Standards and Technology, NIST), and Germany (Physikalisch-Technische Bundesanstalt, PTB) took part in this comparison, whose measurements started in spring 2018 and ended in summer 2019.

II. Experimental Setups

The different NMIs use different measurement procedures, all of which are based on electro-optic sampling (EOS), employing femtosecond lasers with a center wavelength of 1550 nm and an optical pulse duration ranging between 100 and 250 fs. The basic measurement scheme, which is common to all participants, is briefly described in the following; differences between the NMIs are only indicated due to space limitations.

The femtosecond laser beam is split into optical excitation and sampling beams. The excitation beam excites the 1.0 mm coaxially-connectorized photodiode, which generates an electrical pulse. This pulse is coupled by the wafer probe onto a coplanar waveguide (CPW) fabricated on an electro-optic substrate (BIRMM and PTB: GaAs, NIM and NIST: $LiTaO_3$, note that also the dimensions of the CPW differ). The optical sampling beam is used to reconstruct the repetitive electrical waveform generated by the PD at the reference plane of the CPW. For this purpose, different EOS schemes are employed [4,5].

To calculate the electrical waveform at the PD's coaxial connector from the voltage measured by the EOS system, a change in reference plane is necessary. This requires vector network analysis (VNA) to characterize the reflection coefficients of the PD, the CPW and the wafer probe. For this purpose, NIM and NIST employ a conventional VNA [2,4], while BIRMM and PTB employ laser-based VNA [5].

There are numerous sources of uncertainty in the measurements, which differ between the EOS systems. To propagate these uncertainty contributions to the electrical waveform at the PD's coaxial connector, NIST uses the NIST

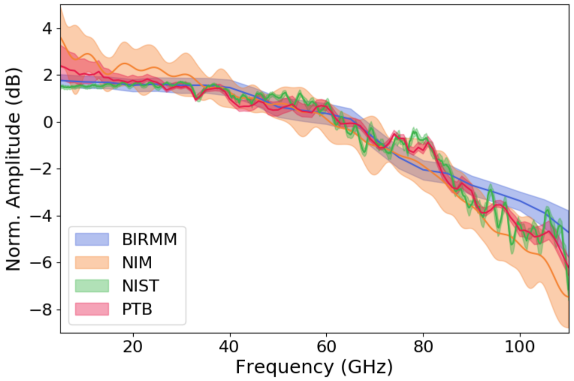


Fig. 1. Frequency-domain amplitude response of the DUT obtained by the different groups. The solid lines and semi-transparent regions denote the best estimates and the coverage intervals, respectively.

[1] Manufacturers name and model are listed here to specify the experimental configuration and do not imply an endorsement. Other makes or model may work as well or better.

Microwave Uncertainty Framework [2], BIRMM, NIM, and PTB use Monte-Carlo simulations [3].

IV. DISCUSSION AND CONCLUSIONS

In all measurements, the PD was biased with 2 V and the optical power was adjusted such that each excitation pulse created a charge of approximately $4 \cdot 10^{-14}$ C. To account for the different pulse durations of the excitation pulses, each NMI specified the photodiode response for a delta-pulse excitation.

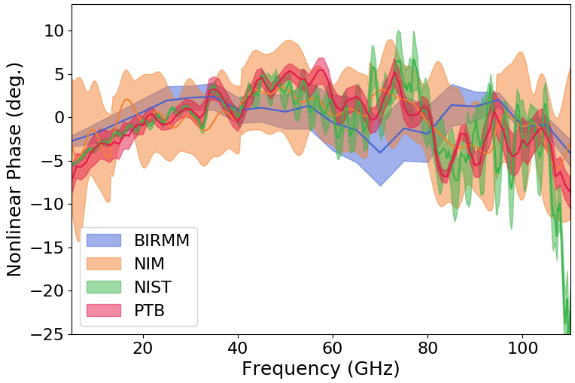


Fig. 2. Frequency-domain phase response of the DUT obtained by the different groups. The solid lines and semi-transparent regions denote the best estimates and the coverage intervals, respectively.

Figure 1 shows the amplitude of the frequency domain responses of the different NMIs. The amplitude spectra were normalized to the average amplitude values in the range from 10 GHz to 100 GHz. It should be noted that differences in time record length between the various NMIs and the choice to interpolate data results in a variety of effects on the presented data including differing frequency step (resolution) and overall smoothness. Furthermore, zero-padding the time record outside of the window around the main pulse can result in a non-physical result by missing any reflections (energy) in the measurement system. An overall good agreement exists with an increased spread between the different results below 20 GHz and above 90 GHz.

The frequency-domain phase responses are shown in Fig. 2. A linear contribution (obtained from linear regression of the phase values in the range from 10 GHz to 100 GHz) was subtracted from the phase values beforehand. Overall, there is very good agreement on the phase with minor difference starting to present above 70 GHz.

Finally, the time-domain responses are plotted in Fig. 3. Here, all pulses have been normalized to one with their maxima centered at 0 ps. The NIST data is not presented here because the traditional network analysis techniques only allow for calibration up to 110 GHz, limiting the ability to inverse FFT data with the same higher frequency content. Similar to the frequency-domain data, there is an overall good agreement with small differences in rise and fall times of the main pulse as well as small discrepancies in the features seen after the main pulse.

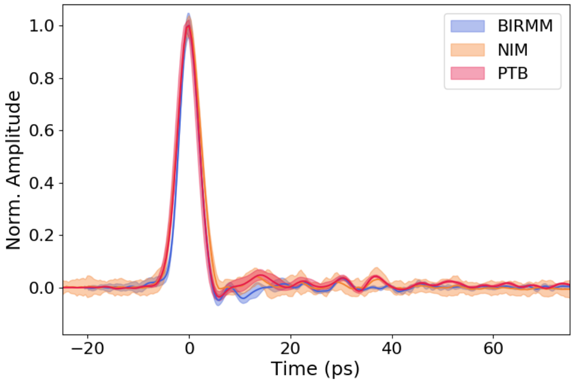


Fig. 3. Time-domain response of the DUT obtained by the different groups. The solid lines and semi-transparent regions denote the best estimates and the coverage intervals, respectively.

Methods for comparing time-dependent measurements are still in their infancy [6] and a more detailed statistical analysis is beyond the scope of this abstract.

In summary, the results of this first comparison look promising, although some features in the frequency- and time-domain responses were not reproduced by all NMIs. Yet, the topic of ultrafast waveform metrology is still in an early stage and additional comparisons will be made to establish better confidence in such measurements.

ACKNOWLEDGEMENT

PTB gratefully acknowledges support by the DFG Research Training Group GrK1952/1.

REFERENCES

[1] P. D. Hale, D. F. Williams, A. Dienstfrey, C. M. Wang, J. A. Jargon, D. Humphreys, M. Harper, H. Füser, and M. Bieler, “Traceability of high-speed electrical waveforms at NIST, NPL, and PTB,” *2012 Conference on Precision Electromagnetic Measurements Digest*, pp. 522-523, Washington, DC, Jul. 2012.

[2] D. F. Williams, A. Lewandowski, T. S. Clement, C. M. Wang, P. D. Hale, J. M. Morgan, D. A. Keenan, and A. Dienstfrey, “Covariance-based uncertainty analysis of the NIST electrooptic sampling system,” IEEE Trans. Microwave Theory Tech., vol. 54, no. 3, pp. 481-491, Jan. 2006.

[3] P. Struszewski, K. Pierz, and M. Bieler, “Time-Domain Characterization of High-Speed Photodetectors,” J Infrared Milli Terahz Waves, vol. 38, no. 11, pp. 1416–1431, Nov. 2017.

[4] D. F. Williams, P. D. Hale, T. S. Clement, and J. M. Morgan, “Calibrating electro-optic sampling systems,” IMS Conference Digest, pp. 1527-1530, May 2001.

[5] M. Bieler, H. Füser, and K. Pierz, “Time-Domain Optoelectronic Vector Network Analysis on Coplanar Waveguides,” IEEE Transactions on Microwave Theory and Techniques, vol. 63, no. 11, pp. 3775–3784, Nov. 2015.

[6] S. Eichstädt, V. Wilkens, A. Dienstfrey, P. Hale, B. Hughs, and C. Jarvis, “On challenges in the uncertainty evaluation for time-dependent measurements,” *Metrologia*, **53** (4), S125-S135, 2016.

Photodiode Calibration Comparison between Electro-Optic Sampling and Heterodyne Measurements up to 75 GHz

Ari Feldman, Jeffrey Jargon, Tasshi Dennis, and Paul Hale
National Institute of Standards and Technology, Boulder, CO USA
ari.feldman@nist.gov

***Abstract* — We present the comparison of a photodiode's measured frequency response calibrated with an electro-optic sampling system and a heterodyne system up to 75 GHz, along with the systems' respective 95% confidence intervals. A brief description of each system and its known sources of uncertainties are provided. The two systems agree to within their respective uncertainties at most frequencies.**

***Index Terms* — Calibration, comparison, electro-optic sampling, heterodyne, measurement, photodiode.**

I. Introduction

The electro-optic sampling (EOS) system at NIST [1, 2] is the United States' primary standard for high-speed waveform calibration and is traceable to the SI through fundamental physics. A photodiode calibrated using this system serves as a time- and frequency-domain transfer standard and allows for subsequent calibrations of high-speed oscilloscopes, electronic comb generators, and high-speed modulated signals [3]. Check-standard photodiodes are frequently measured to verify self-consistent operation of the EOS system, however an external comparison to an independently traceable measurement can help verify the measurement.

At NIST, an independently traceable heterodyne measurement system has been used to measure the magnitude response of photodiodes up to 50 GHz [4, 5]. The beat between two single-frequency lasers defines the excitation of the photodiode. While the EOS system is capable of measuring frequencies up to 110 GHz, the low frequency response is limited below 600 MHz, whereas the heterodyne system provides measurements approaching DC.

Each measurement system has its own set of challenges transforming measured quantities to a calibrated response. The EOS measurement requires that (1) the photodiode operate in the linear regime for excitation with pulsed light; (2) the on-wafer coplanar resistor structures be well characterized; (3) the laboratory environment be stable during measurements; and (4) the probe be de-embedded from the measurement. The heterodyne system relies on a set of calibrated, low-power, diode-based RF power meters to cover the measurement frequency range. For both measurements, calibrated scattering-parameters of the photodiode and connecting devices must be measured and propagated through the photodiode calibration [3].

A previous comparison was presented in [4, 5] but was limited to 50 GHz and didn't include measurement uncertainty. To validate the EOS closer to the frequency of 110 GHz, here we present the comparison of a photodiode calibrated with the EOS and heterodyne measurements up to 75 GHz [6], the highest frequency to date performed at NIST, along with their respective uncertainty estimates. A similar comparison was performed at NPL [7]. In the following sections, we provide a brief description of each system and its major sources of uncertainties, and results of the measurement comparison.

II. Electro-Optic Sampling System

NIST's EOS system is comprised of a 1550 nm mode-locked, erbium-doped fiber laser that emits a series of short optical pulses on the order of 100 fs in duration [1, 2, 3] with a 10 MHz repetition rate. The linearly-polarized output is split into optical excitation and sampling beams. The excitation beam excites the 1.0 mm coaxially-connectorized photodiode, which generates an electrical pulse. This electrical pulse is coupled by the wafer probe onto a coplanar waveguide (CPW) fabricated on an electro-optic LiTaO3 substrate. The optical sampling beam reconstructs the repetitive electrical waveform generated by the photodiode at the on-wafer reference plane in the CPW. This is accomplished by passing the sampling beam through a variable optical delay, polarizing it, and passing it through one of the gaps of the terminated CPW. Since the substrate is electro-optic, the electric field between the CPW conductors changes the birefringence of the crystal, altering the polarization of the optical sampling beam passing through it. A polarizing beam splitter and balanced photoreceiver detects this change, which is proportional to the electric field in the CPW at the instant the optical pulse arrives. This process does not perturb the electrical signal on the CPW. Sweeping the relative delay of the sampling beam allows us to map the voltage at the reference plane in the CPW as a function of time.

To calculate the electrical waveform at the photodiode's coaxial connector from the voltage measured in the CPW by the EOS system, a change in reference plane is required. This requires a VNA to characterize the reflection coefficients of the photodiode and an on-wafer resistor as well as the scattering-parameters (*S*-parameters) of the probe head. The *S*-parameters are then used to determine the impedance levels in the measurement system and determine the frequency-domain voltage the photodiode would generate across a 50 Ω load. The frequency-domain results can then be Fourier-transformed to traceably characterize temporal- and frequency-domain instruments up to 110 GHz, the single-mode limit of the 1.0 mm coaxial connectors of the system.

Feldman, Ari; Jargon, Jeffrey; Dennis, Tasshi; Hale, Paul D. "Photodiode Calibration Comparison between Electro-Optic Sampling and Heterodyne Measurements up to 75 GHz." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020.

There are numerous sources of uncertainty in the EOS system beyond the impedance corrections, including optical reflections from the surfaces of the LiTaO3 wafer, field penetration into the LiTaO3 wafer, the radius of the optical sampling beam, finite temporal widths of the optical sampling and excitation pulses, and measurement repeatability. Correlated uncertainty is propagated through the chain of calibrations using the NIST Microwave Uncertainty Framework [8].

III. Heterodyne System

The NIST heterodyne system for measuring the magnitude of the frequency response of photodiodes operates at frequencies up to 75 GHz. Two single-mode fiber lasers having up to 1 nm of continuous wavelength tuning and a linewidth specification of <1 kHz are combined with free-space optics. The use of free-space optics rather than fiber components reduces the wavelength and polarization dependence of the coupled beams.

After passing through polarizing isolators, fiber-pigtailed collimators are used to collect and deliver the polarized light to the fiber-coupled photodiode under test and the photoreceiver used for frequency measurement. Polarizing and mode-matching the beams in a single-mode fiber is critical to ensuring that the modulation of the light incident on the photodiode is precisely calculable. Adjustable rotary attenuators are used to equalize the power delivered by each laser, as monitored by the photodiode bias current. Tuning of the heterodyne beat frequency between the lasers is performed by controlling the operating temperature of the lasers [4].

The frequency of the heterodyne signal up to 50 GHz is detected by use of an amplified photoreceiver and electrical spectrum analyzer. Above 50 GHz the operating wavelength of each laser is monitored with interferometer-based wavelength meters and the difference frequency is determined.

Aside from impedance corrections, the major sources of uncertainty in the heterodyne system include power meter range scaling, bias current measurement, power sensor noise, and measurement repeatability.

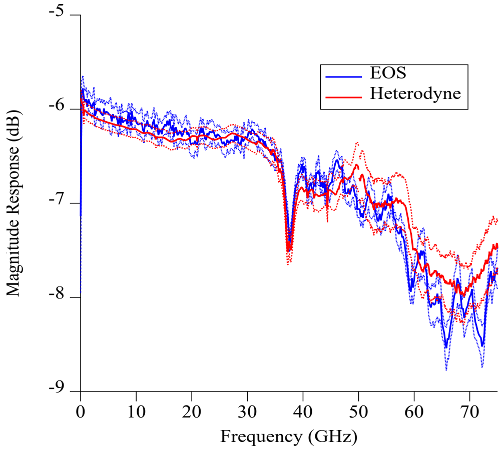


Fig. 1. Comparison of EOS and heterodyne measurements.

IV. Measurement Comparison

The magnitude of the frequency response of a 110 GHz photodiode as measured by the EOS and heterodyne techniques is presented in Fig. 1. The two methods correspond closely in defining the roll-off of the response, with most points agreeing to within their uncertainties. The points that do not agree could be attributed to repeatability of the on-wafer and coaxial connections and calibrations. The EOS measurement indicates more structure despite having a much coarser frequency resolution (200 MHz) relative to the heterodyne measurement (30 MHz). We chose not to connect the two heterodyne curves at 50 GHz, which resulted from the two methods of measuring the beat frequency and different RF power sensors.

The results presented is our best comparison of the methods so far and suggests areas for making meaningful improvements. For the EOS method, improving the frequency resolution from increased range and precision of the delay sweep could reduce distortions in the measured response. The EOS method relies on impedance corrections for the wafer, probe, and photodiode which introduce additional components of uncertainty. The value of the heterodyne method comes from its simplicity and known modulation but relies heavily on the calibrated RF power sensor efficiencies as well as the impedance corrections to isolate the response of the photodiode. The maximum frequency of the heterodyne measurement is limited by the sensitivity of available power sensors to 75 GHz.

V. Conclusion

We presented a comparison of a photodiode's frequency response measured with both EOS and heterodyne methods up to 75 GHz, along with their estimated uncertainties. The results compare favorably while indicating areas for improvement which could result in better agreement.

References

[1] D. F. Williams, et al., "Calibrating electro-optic sampling systems," *IMS Conference Digest*, pp. 1527-1530, May 2001.

[2] P. D. Hale, et al., "Traceability of high-speed electrical waveforms at NIST, NPL, and PTB," *2012 Conference on Precision Electromagnetic Measurements Digest*, pp. 522-523, Washington, DC, Jul. 2012.

[3] P. D. Hale, et al., "Waveform metrology: signal measurements in a modulated world," *Metrologia*, **55** (5), S135-S151, 2018.

[4] T. Dennis and P. D. Hale, "High-accuracy photoreceiver frequency response measurements at 1550 mm by use of a heterodyne phase-locked loop," *Optics Express*, vol. 19, no. 21, pp. 20103-20114, 2011.

[5] T. S. Clement, et al., "Calibrated photoreceiver response to 110 GHz," Conference Digest of the15th annual meeting of the IEEE Lasers and Electro-optics Society, Nov. 10-14, 2002, Glasgow, Scotland.

[6] A. Feldman, et al., "Photodiode Calibration Comparison between Electro-optic Sampling and Heterodyne Measurements up to 75 GHz" 32nd URSI GASS, Montreal, 19–26 August 2017.

[7] D. A. Humphreys, et al., "Vector Calibration of Optical Reference Receivers Using a Frequency-Domain Method," IEEE Trans IM, Vol. 54, No. 2, pp. 894-897, April 2005.

[8] D. F. Williams, et al., "Covariance-based uncertainty analysis of the NIST electrooptic sampling system," IEEE Trans. Microwave Theory Tech., vol. 54, no. 3, pp. 481-491, Jan. 2006.

Feldman, Ari; Jargon, Jeffrey; Dennis, Tasshi; Hale, Paul D. "Photodiode Calibration Comparison between Electro-Optic Sampling and Heterodyne Measurements up to 75 GHz." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020.

Graphene quantum Hall effect devices for AC and DC resistance metrology

Mattias Kruskopf[*], Dinesh K. Patel[§,†], Chieh-I Liu[§,‡], Albert F. Rigosi[§], Randolph E. Elmquist[§], Yicheng Wang[§], Stephan Bauer[*], Yefei Yin[*], Klaus Pierz[*], Eckard Pesel[*], Martin Götz[*] and Jürgen Schurr[*]

[*]Physikalisch-Technische Bundesanstalt (PTB), Bundesallee 100, Braunschweig, 38116, Germany

[§]National Institute of Standards and Technology (NIST), 100 Bureau Drive, Stop 8171, Gaithersburg, MD, 20899, USA

[†]Graduate Institute of Applied Physics, National Taiwan University, Taipei 10617, Taiwan

[‡]University of Maryland, Department of Chemistry and Biochemistry, College Park, MD, 20742, USA

Abstract — **The frequency dependence of the quantized Hall resistance at alternating current results from capacitive losses inside the sample as well as between the sample and external parts. In this joint effort, we report on ac quantum Hall measurements of a graphene-based Hall bar using superconducting contacts and a novel contact design approach.**

Index Terms **— electrical measurement standards, ac quantized Hall resistance, epitaxial graphene.**

I. Introduction

The fundamental nature of the quantum Hall effect allows for the realization of an electrical standard for the direct current (dc) resistance as well as for the impedance, capacitance, and inductance [1]–[3]. To accelerate the worldwide adoption of these techniques, National metrology institutes are investigating the deviation of the quantized Hall resistance (QHR) at alternating current (ac) from the value measured at dc. Parasitic capacitances were identified to be the reason for the observed offsets, which can be compensated using externally applied shields [4].

In this paper, we report on a joint effort of the PTB and the NIST to characterize an epitaxial graphene-based Hall bar using superconducting contacts and an optimized contact geometry to identify measurement conditions and sample characteristics that are critical to the measurement precision [5].

II. DC Measurements

Figure 1(a) shows the sample design, the measurement configuration and dominating parallel capacitances under ac conditions. The magnetotransport properties at direct current were pre-characterized using a current source and a nano voltmeter for magnetic flux densities between $B = 0$ T to $B = 12$ T. Figure 1(b) shows that due to the low carrier density of $n = 6.6\times10^{10}$ cm^{-2}, the resistance plateau started at around $B = 1$ T and the longitudinal resistance, measured with current reversal to reject thermal voltages, was within the setups the noise floor for $B > 2$ T. The sample was then characterized with a cryogenic current comparator at $B = 7$ T. The longitudinal resistance was found to be (465 ± 1.5) µΩ on the low potential side and (694 ± 1.5) µΩ on the high potential side (type A uncertainty $k = 1$).

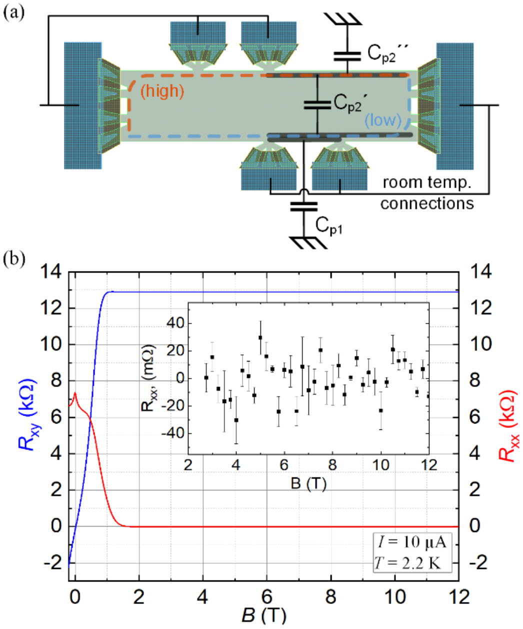


Fig. 1. (a) The drawing shows the sample design as well as indicates the origin of parallel capacitances of the Hall and longitudinal resistance measurements at ac. (b) Characterization of dc magneto transport properties. The carrier density n and mobility μ are 6.6×10^{10} cm^{-2} and 12883 cm²/Vs, respectively. The inset indicates a vanishing of the longitudinal resistance at magnetic flux densities around $B = 2$ T. The error bars represent the standard deviation determined from three measurements.

III. AC Measurements

The deviation of the Hall resistance at ac from the dc value R_H shows a linear frequency dependence with a slope of (81.7 ± 1.5) × 10^{-9} / kHz (see Fig.2, top). The parallel capacitance C_{p2} (bottom) is measured at the pair of Hall contacts and is

Kruskopf, Mattias; Patel, Dinesh; Liu, Chieh-I; Rigosi, Albert; Elmquist, Randolph; Wang, Yicheng; Bauer, Stefan; Yin, Yefei; Pierz, Klaus; Pesel, Eckard; Goetz, Martin; Schurr, Jurgen. "Graphene quantum Hall effect devices for AC and DC resistance metrology." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020.

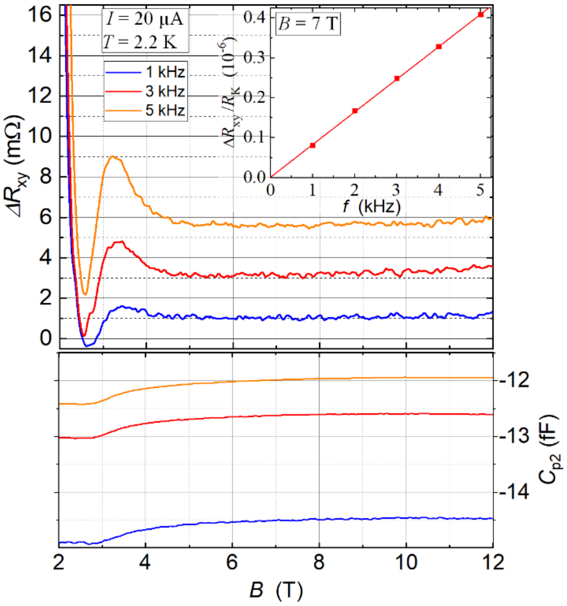


Fig. 2. The deviation of the Hall resistance ΔR_{xy} at ac from the dc value R_H (top) with a slope of 81.7 ± 1.5 × 10^{-9} /kHz (see inset) and corresponding magetocapacitance measurements (bottom).

composed of the two components C_{p2}' and C_{p2}'' where C_{p2} = C_{p2}'- C_{p2}''. While the absolute value of C_{p2} has a large uncertainty of 58 fF, the shape of the magnetic field dependence can still be precisely measured. The graph shows that C_{p2} increases in the beginning of the resistance plateau and becomes mostly flat for $B \geq$ 6T, reflecting the changes in the areas of compressible and incompressible states inside the QHR sample and a dominating contribution from C_{p2}'' compared to C_{p2}'. Figure 3 shows the frequency dependence of the longitudinal resistance (top) and the parallel capacitance C_{p1} (bottom) measured at the same contact pair. The upwards bending of the longitudinal resistance curves at higher magnetic flux densities and the absence of a clear flat region indicates a non-zero longitudinal resistance at f = 0 Hz in agreement with the results of the cryogenic current comparator. Interestingly, the increase of the longitudinal resistance at higher magnetic fields seems to be associated with an increase of C_{p1}.

IV. Conclusion and Outlook

The applied sample design resulted in a Hall resistance with a positive frequency dependence. Previously published ac QHR measurements using graphene Hall bars [6] showed a positive frequency dependence in the case of much larger sample dimensions (e.g. 2600 μm × 800 μm) and higher current. By further investigating the frequency dependence with respect to the charge carrier density and the applied current as well as by applying the double-shield technique, we are aiming to find favorable conditions for precise ac QHR measurements.

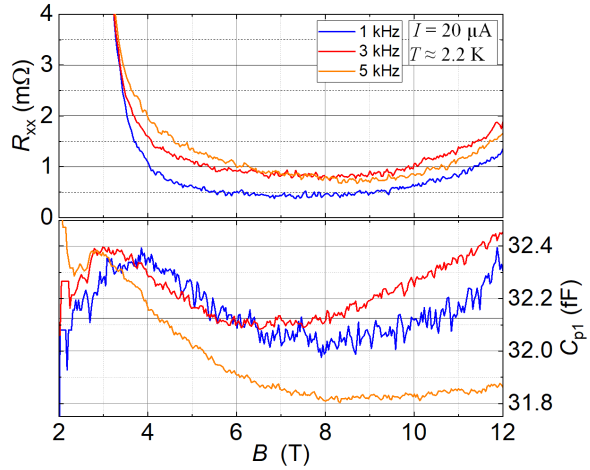


Fig. 3. Longitudinal resistance (top) and magnetocapacitance measurements (bottom) at ac and different frequencies at the low potential side.

Acknowledgment and Funding

This work was supported by the Joint Research Project GIQS (18SIB07). This project received funding from the European Metrology Program for Innovation and Research (EMPIR) co-financed by the Participating States and from the European Unions' Horizon 2020 research and innovation program. The work of D. K. Patel at NIST was made possible by arrangement with Prof. C.-T. Liang of the National Taiwan University.

References

[1] J. Schurr, V. Bürkel, and B. P. Kibble, "Realizing the farad from two ac quantum Hall resistances," *Metrologia*, vol. 46, no. 6, pp. 619–628, Dec. 2009.

[2] J. Schurr, F. Ahlers, and B. P. Kibble, "The ac quantum Hall resistance as an electrical impedance standard and its role in the SI," *Meas. Sci. Technol.*, vol. 23, no. 12, p. 124009, Dec. 2012.

[3] A. Hartland, "The Quantum Hall Effect and Resistance Standards," *Metrologia*, vol. 29, no. 2, pp. 175–190, Jan. 1992.

[4] J. Schurr, B. P. Kibble, G. Hein, and K. Pierz, "Controlling losses with gates and shields to perfect a quantum Hall impedance standard," *IEEE Trans. Instrum. Meas.*, vol. 58, no. 4, pp. 973–979, 2009.

[5] M. Kruskopf *et al.*, "Two-Terminal and Multi-Terminal Designs for Next-Generation Quantized Hall Resistance Standards: Contact Material and Geometry," *IEEE Trans. Electron Devices*, pp. 1–5, 2019.

[6] C.-C. Kalmbach, *AC-Quanten-Hall-Effekt und 1/f -Rauschen in epitaktischem Graphen*, E 113. PTB Bericht, 2017.

Kruskopf, Mattias; Patel, Dinesh; Liu, Chieh-I; Rigosi, Albert; Elmquist, Randolph; Wang, Yicheng; Bauer, Stefan; Yin, Yefei; Pierz, Klaus; Pesel, Eckard; Goetz, Martin; Schurr, Jurgen. "Graphene quantum Hall effect devices for AC and DC resistance metrology." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020.

SP-1133

POWDER THERMAL CONDUCTIVITY MEASUREMENTS IN L-PBF USING POWDER-INCLUDED BUILD SPECIMENS: INTERNAL GEOMETRY EFFECT

Shanshan Zhang[a, 1], Brandon Lane[b], Kevin Chou[a]
[a]Department of Industrial Engineering
University of Louisville, Louisville, KY, 40292
[b]Engineering Laboratory, National Institute of Standards and Technology
Gaithersburg, MD, 20899

ABSTRACT

This study investigates the thermal conductivity of 17-4PH stainless steel powder that was encapsulated within specimens with different internal geometries in laser powder bed fusion (L-PBF) additive manufacturing (AM). The objective is to evaluate the effect of the internal geometry of the specimens on the measurement of the powder thermal conductivity and to compare the thermal properties amongst the 17-4PH and two additional powder materials used in L-PBF. Continued from the previous work [1], three new cone configurations in the hollow specimens were designed and fabricated in an L-PBF system. The thermal conductivity of the internal powder was indirectly measured using an experimental-numerical approach, combined with laser-flash testing, finite element (FE) heat transfer modeling and multivariate inverse method. The results reveal that the thermal conductivity of 17-4PH powder ranges from 0.67 W/(m·K) to 1.34 W/(m·K) at 100 °C to 500 °C, and varies with the internal geometry of the specimens. In addition, the measurement of the hollow specimen with a convex cone seems to be a more reliable evaluation. Further, the thermal conductivity ratio of the powder to the solid counterpart of 17-4PH approximately ranges from 3.9 % to 5.5 % at tested temperatures, which is similar to the results obtained from the nickel-based super alloy 625 (IN625) and Ti-6Al-4V (Ti64) powders measured in a previous study.

Keywords: Laser powder-bed fusion; laser flash; finite element modeling; inverse method; powder thermal conductivity; 17-4PH stainless steel

1. INTRODUCTION

Powder bed fusion (PBF) additive manufacturing (AM) is ever-increasingly investigated and widely adopted in recent years because of its capability to fabricate solid freeform shaped parts and high quality products for a wide range of applications. During the building process, metallic powder is spread onto the building plate and successive layers of powder are selectively fused by a high-energy heat source [2]. The thermal transfer between the melted part and the surrounding powder bed have a profound impact on the temperature gradient and the solidification rate, which in turn, gives rise to a substantial influence on the grain growth and microstructural morphology [3-6] and resultant mechanical properties of the final part [7, 8]. In addition, the powder bed acts as a support for overhanging structures in PBF, but also acts as a thermal insulator. This creates localized overheating, dependent on the relative amount of solid compared to powder material near the melt pool [9, 10]. Owing to the high cost of PBF equipment and materials, researchers seek assistance from the computational approaches to understand the thermal behavior in PBF processes, and consequentially to improve the fabrication processes and part quality. A key input for building these simulations is the thermal properties of the powder bed, which are essential to achieve reliable computational predictions.

Spread by the recoater and infiltrated by an inert gas environment in PBF, the powder bed can be regarded as a mixture of gas-infiltrated particles with a specific packing density. The individual properties of the metal powder, gas, and packing density can be combined to form an assumed continuum or singular material. For decades, many have recognized the dominant role of conduction in the heat transfer in heterogeneous gas-solid systems in different types of application processes, such as chemical reactors [11, 12], drying systems [13] and heat exchangers [14]. It has been reported that many factors can be critical, such as volume fraction, contacts between particles, infiltrated gas types and gas pressure [15-22]. For example, Yagi et al. [18] established a model of heat transfer in packed bed with

[1] Contact author: shanshan.zhang@louisville.edu

Zhang, Shanshan; Lane, Brandon; Chou, Kevin. "Powder Thermal Conductivity Measurements in L-PBF using Powder-Included Build Specimens: Internal Geometry Effect." Paper presented at Manufacturing Science and Engineering Conference (MSEC2020), Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

motionless gas. They raised theoretical equations to predict the effective thermal conductivities of a gas-powder packed bed system and concluded that the radiant heat transfer is more effective when the temperature is higher than 400 °C. Wakao and Vortmeyer [20] claimed that the effective thermal conductivity of packed beds with stagnant gas is primarily affected by the conductivity of the packed bed and radiation in the gas-solid system, as well as the contact conductivity between particles. A discrete element method in granular material heat transfer was developed by Vargas and McCarthy [22], which considered the effect of stress and contact heterogeneities on the pressure distribution in the stacked particles. On the other hand, Wei et al. [23] and Bala et al. [24] presented an experimental method to measure the thermal conductivities of metallic powders, and the latter pointed out that the gas infiltrating has an effect to the heat dissipation of the powder bed.

Recent studies by Cheng et al. [25] and Zhang et al. [1, 26] estimated the thermal conductivity of metallic powder using a hybrid laser flash experiment and numerical heat transfer simulation. In that work, the authors designed a hollow specimen to encapsulate powder during fabrication to maintain the in-situ powder bed conditions. Using an inverse heat transfer approach, the thermal properties of the encapsulated powder could be extracted from the laser flash measurement results. Herein, except for the thermal conductivity and powder porosity, the contact conductance between powder and the specimen shell was also critical to the heat transfer analyses and evaluated as an output. Additionally, a critical finding in those studies was that the powder thermal conductivity obtained by this method seems to be varied for different internal geometries of specimens. It was noted that a gap exist between the internal powder and the top shell of the specimen, which adversely affects heat flow through the specimen. Although this finding provides an insight of measurement uncertainty potentially related to the specimen geometry, quantitative understanding of the specimen internal geometry effect on the powder thermal property analysis is still lacking.

To continue the previous work, the objective is to design addition specimen geometries to investigate their effects on the measurement of powder thermal properties. In addition, the contact conductance between the powder and internal surfaces of specimen shell with different geometries are of interest. Furthermore, the experimental-numerical approach is extended to investigate the temperature-dependent thermal conductivity of 17-4PH stainless steel powder. A comparison of powder thermal conductivity and porosity with two additional powder materials: nickel-based super alloy 625 (IN625) and Ti-6Al-4V (Ti64), presents different results in a tested temperature range of 100 °C to 500 °C.

2. EXPERIMENTAL DETAILS

The test specimens were designed hollow discs that enclosed powder, thus maintaining the as-fabricated powder conditions. As indicated in [25], a cone feature can reduce the gap and improve the contact condition between the powder and internal upper shell, and therefore, increase the accuracy of simulation. However, the follow-up investigation in [1] showed inconsistent measured thermal conductivity of metallic powder, including IN625 and Ti64 materials, at specimens with three different internal geometries. To clarify the effect of the internal geometry of the specimens, the three shaped internal geometries were also tested in this study for estimating the powder properties of 17-4PH powder. In addition, 3 developed geometric features were designed as shown in Figure 1, including (1) top cone with a height of 1 mm (1Cone-1.0), (2) spherical cap with a height of 1 mm (Convex), and (3) concave, rotated cap with the height of 1 mm (Concave). These three novel specimens only vary the top internal geometry and keep the same bottom geometry to test the contact conditions between the powder and the top shell, specifically.

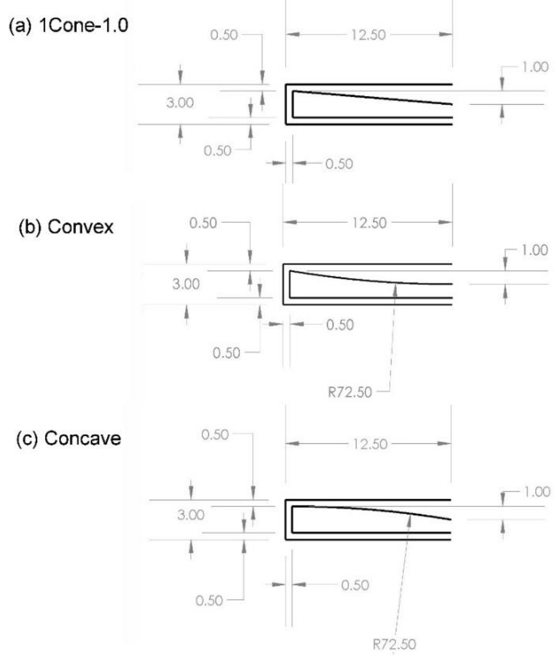


FIGURE 1: SYMMETRICAL CROSS-SECTION VIEW OF DIFFERENT SPECIMEN DESIGNS SHOWING INTERNAL GEOMETRY (UNIT: MM)

In this study, 17-4PH stainless steel supplied by LPW Technology[2] was used for specimen fabrication by an EOSINT M270 L-PBF system. The build orientation for the specimens

[2] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

was along the diameter direction (along the horizontal axis in Figure 1) to reduce the necessity of support structure. The machine recommended process parameters for 17-4PH solids were used, resulting in a scan speed of 1000 mm/s, a laser power of 195 W, hatch spacing of 100 µm, and a layer thickness of 20 µm. There was no laser exposure within the internal hollow region to encapsulate loose powder. In addition, a N_2 inert gas environment was employed during fabrication.

A DLF-1200 laser flash system from TA Instruments was utilized to acquire thermograms from each sample, which indicate the time-dependent temperature rise measured via pyrometer on one side of the sample resulting from an applied laser pulse on the other. Thermal diffusivity values, representing the homogeneous properties of the entire sample, are extracted using Clark-Taylor method [27] in this study. Before testing, a graphite coating was uniformly sprayed on the sample surface to maximize the absorption of the laser pulse. The sample was then put into the chamber sitting in a sample holder. During testing, the ambient temperature in the chamber, which also sets the sample temperature, was a user-defined preset. When the environment and sample temperatures reach steady-state and equilibrium, a 3 ms laser pulse of 25 J was applied to the bottom of the specimen over a uniformly distributed, circular region of 22 mm diameter. An infrared pyrometer received the voltage signal from the top of the specimen from a round region with a diameter of about 9.6 mm for 60 s. The time-dependent voltage signal, called a thermogram, was then obtained and normalized for use in the inverse-heat transfer simulation. Three samples for each type of internal geometry and three laser pulses for each sample were conducted in the laser flash system for a repeatability investigation at each preset temperature. The tested temperatures ranged from 25 °C to 500 °C.

3. Test sample system and FE model

Test samples used in the FE model were dimensionally equal to the real test components, which were assumed homogeneous materials in the modeling. The mesh for the sample system consists of 10-node quadratic heat transfer tetrahedron bricks, with the mesh sizes of 0.5 mm for the specimen and 0.7 mm for the holder, respectively. The total heat flux applied to the bottom surface of the specimen was simplified as a uniform distribution. The heat transfer takes place through three mechanisms: (1) heat conduction in the specimen, (2) heat conduction between the specimen and the holder, and (3) heat loss due to convection and thermal radiation from the sample system to the surrounding environment. The ambient temperature in the modeling was set the same as the actual experimental environment. Further details on the FE model construction are provided in [1].

To calculate the thermal properties of the internal powder, there existed five unknowns in the FE model: (1) the specimen-holder contact conductance (k_p), (2) powder density (ρ), (3) powder thermal conductivity (k), (4) the contact conductance between the powder and the top solid shell (k_t), and (5) the contact conductance between the powder and the bottom solid shell (k_b). Two major steps were conducted to estimate the five unknowns. First, the solid sample system was simulated and measured to analyze the k_p, in order to simplify the model and reduce the computational workload. Then, the k_p in the solid sample system was assumed the same values in the powder-enclosed sample system at each corresponding temperature, and utilized as a known parameter in the latter model. Next, the remaining four unknowns related to the powder were estimated using an experimental-numerical inverse heat transfer method, originally derived in [28], and applied to the laser flash system as described in [25]. To optimize the estimated powder properties, the multivariate inverse method was adopted to evaluate a set of unknown properties in each iteration and reduce the difference between the simulation and experimental results.

4. RESULTS AND DISCUSSION

4.1 Laser flash experimental results

The specimens with different internal geometries were measured in the laser flash system at environment temperature range of 25 °C to 500 °C. Figure 2 shows the effective homogeneous sample thermal diffusivity on the three new designed specimen geometries, which are obtained from the laser flash instrument using the Clark-Taylor analytical model. The resulting diffusivity ascends with the rising temperature from 25 °C to 300 °C, and then a plateau occurs when the temperature continues increasing until 400 °C, and then a slight increase at 500 °C. The values range from 0.0018 cm^2/s to 0.0025 cm^2/s at 25 °C. In addition, at each temperature, the Convex specimens provide highest thermal diffusivity among the three types of specimens due to the largest solid to powder mass ratio, followed by 1Cone-1.0 and Concave, successively.

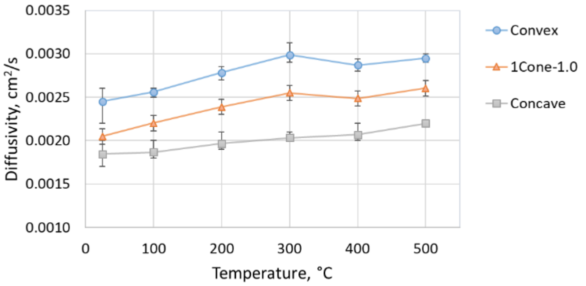


FIGURE 2: COMPARISON OF HOMOGENEOUS TOTAL SAMPLE THERMAL DIFFUSIVITY OF SAMPLES WITH DIFFERENT INTERNAL GEOMETRIES

Furthermore, the time-response thermograms (normalized pyrometer voltage) were obtained from the experiments. Figure 3 (a) shows the thermograms at five tested temperatures of Concave specimens as an example. It can be noted that the plots shift to the left gradually with an increasing temperature; the peak temperature occurs sooner, from 22 s at 100 °C to 16 s at 500 °C, indicating an increase in sample thermal diffusivity. Besides, at a certain temperature, the thermogram of the Convex specimen exhibits a faster heating rate than the other two, and the Concave shows the slowest, corresponding to the thermal diffusivities shown in Figure 2. An example of the comparison between the three cone geometrical specimens at 200 °C is shown in Figure 3 (b).

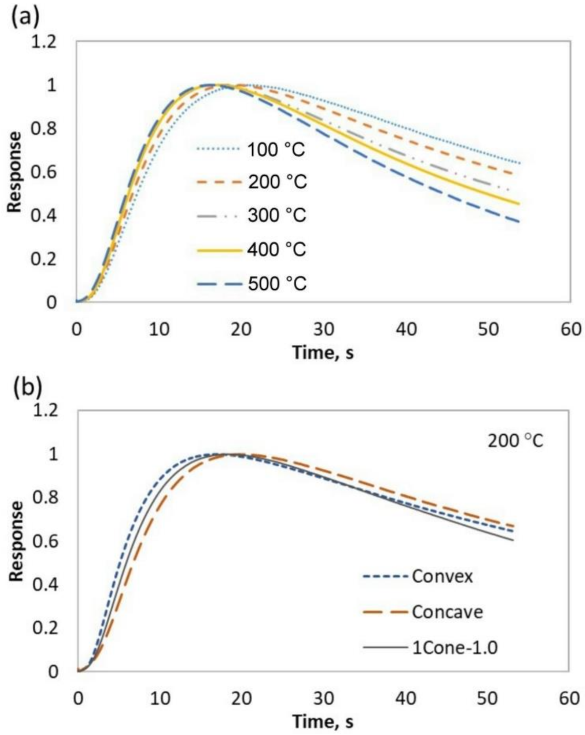


FIGURE 3: EXPERIMENTAL THERMOGRAMS FOR (A) CONCAVE SPECIMENS AT TESTED TEMPERATURES; AND (B) THREE TYPES OF SPECIMENS AT 200 °C

4.2 Simulation results

The FE model was established using measured dimension of the as-built specimens fabricated by L-PBF, and the material properties of the specimens were employed using the information of solid 17-4PH material [29, 30] in Figure 4. Radiation heat loss boundaries assumed an emissivity for 17-4PH of 0.2 [31], and the convection coefficient was estimated as 10 W/(m^2·K) [25]. In addition, the alumina sample holder was included in the modeling and applied the material properties of alumina [32-34].

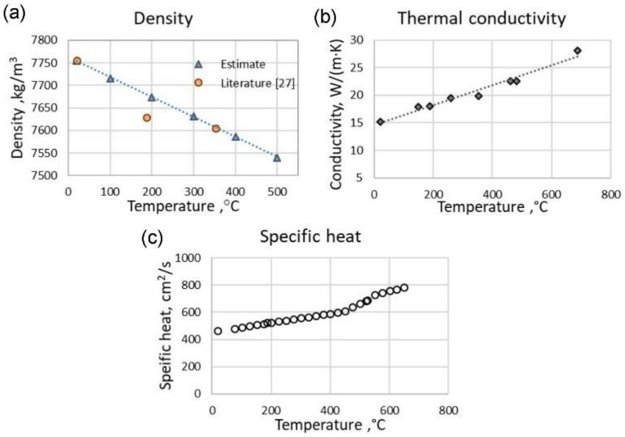


FIGURE 4: SOLID MATERIAL PROPERTIES FOR 17-4PH USED FOR SHELL STRUCTURE IN THE FINITE ELEMENT MODELS [29, 30].

4.2.1 Powder thermal conductivity with different cone geometries

The extracted powder thermal conductivity using the inverse heat transfer method applied to the three new cone designs and is shown in Figure 5. The powder conductivity achieved ranges from 0.65 W/(m·K) to 1.34 W/(m·K), and displays a linear increase with temperature. Additionally, the Convex specimen presents about 0.22 W/(m·K) to 0.31 W/(m·K) higher thermal conductivity than Concave at corresponding temperatures, with the 1Cone-1.0 model located in the middle of the range.

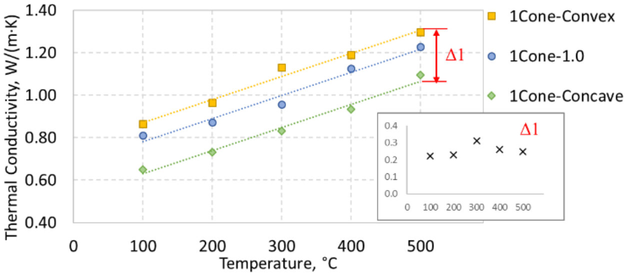


FIGURE 5: POWDER THERMAL CONDUCTIVITY COMPARISON IN 3 NEW CONE GEOMETRIES (INLAYED PLOT USES THE SAME COORNIDATE UNITS WITH THE MAJOR PLOT)

The powder thermal conductivity using the three models in [1] is shown in Figure 6. Similar to the other three models and the two materials in [1], the powder thermal conductivity also exhibits nearly linear to the temperature. Meanwhile, the two models of 2Cone-0.25 (2 cones 0.25 mm thick) and 1Cone-0.5 present similar simulation results that are generally lower than 2Cone-0.5 model. The difference (Δ2) between 2Cone-0.5 specimen and the other two is in a range of 0.10 W/(m·K) to 0.19 W/(m·K). It is indicated that the difference with new cone configurations generally shows 35 % wider (Δ1>Δ2).

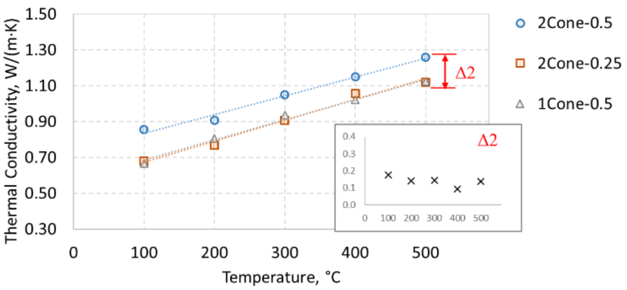


FIGURE 6: POWDER THERMAL CONDUCTIVITY COMPARISON IN 2CONE-0.5, 2CONE-0.25 AND 1CONE-0.5 MODELS (INLAYED PLOT USES THE SAME COORNIDATE UNITS WITH THE MAJOR PLOT)

4.2.2 Analysis of powder conductivity ratio ($\Delta k/k_{ref}$) vs. cone volume ratio ($\Delta V/V_{ref}$) upon the reference model

Stemming from the apparent specimen geometry effect, further investigation on powder thermal conductivity of 17-4PH powder results are analyzed in this section. The cone volume (V) was estimated by taking the difference between the measured mass of the sample, and that of a hypothetical hollow disk of the same external geomery, but 0.5 mm thick skin (and no cone structure). For example, the cone volume of 1Cone-0.5 model was calculated using the cone geometry with a cone height of 0.5 mm. To compare between the different specimens, 1Cone-1.0 specimen was set as the reference (named as “ref”). A measurement criterion was set as a ratio of $\Delta k/k_{ref}$, where k_{ref} is the thermal conductivity of the reference specimen, and Δk equals the objective conductivity value subtracted by that of the reference at the corresponding temperature. Likewise, the volume ratio ($\Delta V/V_{ref}$) is defined based upon the cone volume of the reference. The $\Delta k/k_{ref}$ ratio vs. $\Delta V/V_{ref}$ ratio was plotted in Figure 7. At 100 °C, it is noticed that $\Delta k/k_{ref}$ ratios of 2Cone-0.25 and 1Cone-0.5 are close at about 16% to 17 %. Concave has a distinguishable (2%) lower $\Delta k/k_{ref}$ ratio despite little difference in $\Delta V/V_{ref}$. Additionally, compared to the reference, 2Cone-0.5 shows higher $\Delta k/k_{ref}$ with about 3 % variation. On the other hand, unlike to the three models with smaller cone volume, the Convex model shows about only 6 % $\Delta k/k_{ref}$ above the x-axis, while the cone volume is about 1.5 times of the reference specimen. Similar quantitative analysis at 100 °C shows the unbalanced $\Delta k/k_{ref}$ ratio at both sides of the y-axis although the $\Delta k/k_{ref}$ ratio for smaller cone models exhibits smaller difference than those at 500 °C. Therefore, such a non-linear correlation between $\Delta k/k_{ref}$ ratio, and reduced difference in this ratio with different cone geometries at higher $\Delta V/V_{ref}$ ratio indicates more reliability of the powder conductivity measurement at a higher cone volume level, such as for the Convex specimens.

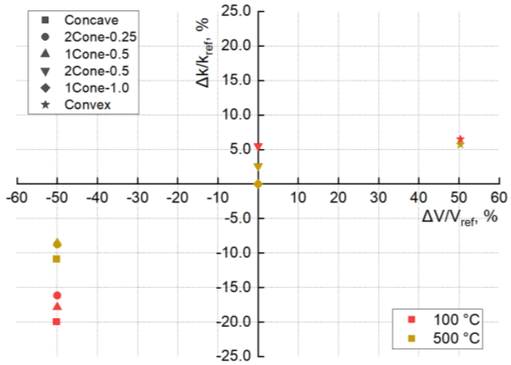


FIGURE 7: $\Delta K/K_{REF}$ VS. CONE VOLUME RATIO AT 100 °C AND 500 °C

4.2.3 Powder thermal conductivity in different materials

Normalized by the solid thermal conductivity (k_s) of the respective powder materials, the powder thermal conductivity ratio (k/k_s) with respect to the fabricated materials in 2Cone-0.5 samples is compared in Figure 8. It can be noticed that the k/k_s ratio of IN625 powder shows a temperature-dependent descent with from 6.9 % to 5.2 %. In contrast, 17-4PH powder exhibits a range of 5.0 % to 5.7 % and a slightly increasing trend with an increasing temperature from 100 °C to 500 °C while the resultant difference between both end temperatures appears insignificant. On the other hand, Ti64 powder keeps an approximately constant k/k_s ratio at all tested temperatures, and the value of the ratio is close to that of 17-4PH powder, which is about 5%.

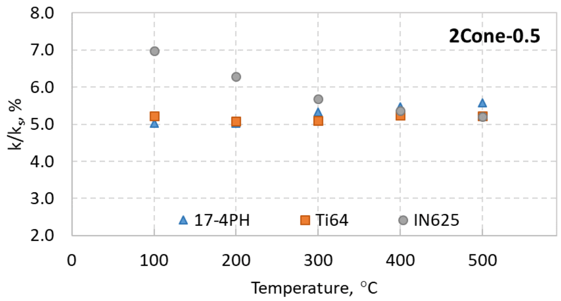


FIGURE 8: COMPARISON OF TEMPERATURE-DEPENDENT K/K_S RATIO IN 2CONE-0.5 MM MODEL WITH THREE MATERIALS.

Similar to 2Cone-0.5 model, IN625 powder still exhibits a descending k/k_s ratio and 17-4PH powder shows a slightly increasing k/k_s in both of 1Cone-0.5 and 2Cone-0.25 models from 100 °C to 500 °C. However, in both models, Ti64 powder shows an approximately 0.5% decreasing k/k_s ratio. In addition, the k/k_s ratio for the models of 1Cone-0.5 and 2Cone-0.25 in the three powder materials generally reduces 1% or so due to the internal cone configuration effect. Figure 9 shows the temperature-dependent k/k_s ratio for 2Cone-0.25 models with three powder materials.

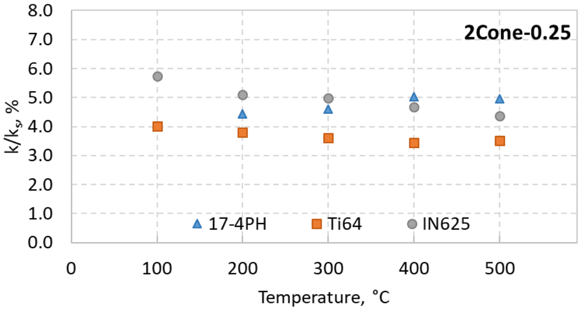


FIGURE 9: COMPARISON OF TEMPERATURE-DEPENDENT K/KS RATIO IN 2CONE-0.25 SPECIMEN WITH THREE MATERIALS

4.2.4 Thermal contact conductance comparison

It has been recognized that the contact conductance is a function of several parameters, such as contacting interface geometry, surface roughness, temperature, interfacial pressure, etc. [35-39]. In the multivariate inverse method, as another two outputs from the simulation, thermal contact conductance between the powder and solid shell on the top and at the bottom are principal parameters interfering with heat transfer in the powder-enclosed specimens. Among the three new specimen geometries, the top and bottom contact conductance are compared in Figure 10. It is noticed that the bottom contact conductance values vary little between the three specimens at each temperature, but overall increase with temperature. The reason for the similar bottom conductance could be that the three specimens have the same bottom geometry. Besides, the bottom contact conductance in all three specimens shows a temperature-dependent increase from 100 °C to 300 °C. Subsequently, there exists a slight retrogression at 400 °C and then a rebound at 500 °C. This trend was also observed for the measured total sample homogeneous thermal diffusivity in Figure 2, indicating this contact conductance may either be a major contributor to the total diffusivity, or may be similarly affected by intrinsic temperature dependence of the material thermal properties.

On the other hand, for the top contact conductance, Convex exhibits apparent higher values than the other two specimens at all tested temperatures; wherein, 1Cone-1.0 shows a higher or similar (at 500 °C) values than Concave. Likewise, at 400 °C, a retrogression or stagnation occurs at 400 °C for all three specimens, followed by an increasing at 500 °C.

Moreover, the contact conductance (top and bottom) were compared between Ti64, IN625 and 17-4PH, and an example of 2Cone-0.25 for the three materials is shown in Figure 11. It can be seen that the bottom contact conductance for all three materials are generally higher than that on the top of the internal powder and solid shell. This is considered to result from a gap that exists between the top shell and the internal powder, interfering the heat transfer through the specimens. Additionally, the retrogression also occurs in 17-4PH 2Cone-0.25 specimen at 400 °C, while such a phenomenon is not found in the other two materials.

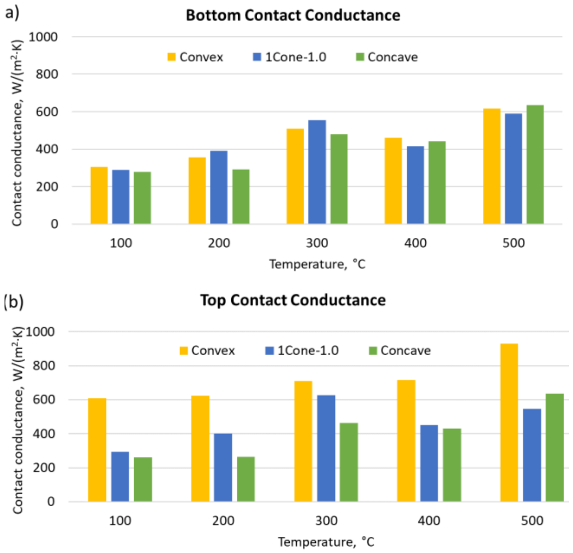


FIGURE 10: SIMULATED CONTACT CONDUCTANCE (A) AT THE BOTTOM, AND (B) ON THE TOP OF THE INTERNAL POWDER AND SOLID SHELL

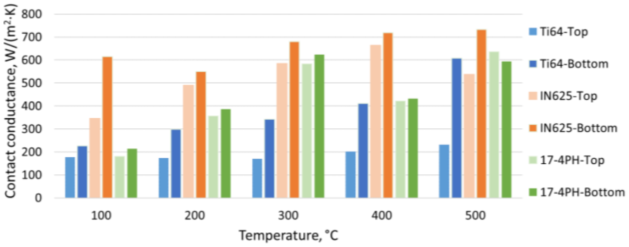


FIGURE 11: CONTACT CONDUCTANCE EVALUATION IN THREE MATERIALS

4.2.5 HEAT TRANSFER IN SPECIMENS

As the transient heat flux is exposed at the bottom of the specimens, the heat flux dissipates through the bottom shell towards the internal powder, and a lower temperature band occurs along the contact surfaces of the three specimens. It is observed that at t = 2.183 s, the internal powder takes heat, spreading upwards slowly due to limited heat transfer at the contact region, with the center region leading higher temperature and dissipation around the center. Additionally, the side shell provides more ability to carry heat to the upper shell, and thus the powder about the edge shows higher temperature for the three specimens. However, apparently, Convex gives a faster thermal flow passing through, followed by 1Cone-1.0 and Concave in an order. At t = 24.863 s, the heat transfer reaches steady state in Convex, and nearly so in Concave and 1Cone-1.0 specimens.

The corresponding heat flux distribution in the middle cut-off areas of the 3 internal powders is shown in Figure 12. At t = 0.011 s, the heat flux starts showing up at the bottom of the internal powders in the three specimens, and not obviously different in values. As the heat is outspreading, heat flux vectors of approximately 6.7 W/m^2 occasionally occurs around the central powder at about 0.1 s in all three specimens. Sequentially, at t = 2.183 s, the energy flows increase in the powder, and the middle portion of the powder shows a higher heat flow rate. Compared to the other two cone featured internal powders, the Concave specimen exhibits a narrower region specifying high heat flux value between 4.0 W/m^2 and 5.0 W/m^2, as the double-arrows are shown in Figure 12.

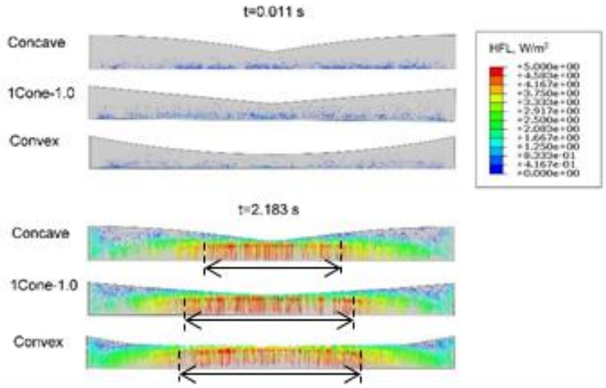


FIGURE 12: HEAT FLUX DISTRIBUTION IN POWDERS IN THREE CONE FEATURED SPECIMENS AT DIFFERENT TIMES

On the other hand, at t = 2.183 s, the heat flux exhibits significantly different on the top of the outer shells with different cone configurations, as shown in Figure 13. In Convex specimen, the heat flux displays a higher value, approximately 2.5 W/m^2 at the boundary that contacts the internal powder, and gradually reduces to none upwards but increases laterally toward outside. The heat flux in the top shells of 1Cone-1.0 and Concave specimens also exhibits the similar transition, but the heat flux at the bottom boundaries is lower than that in Convex. The bottom shells for the three cone featured specimens display insignificant differences. Beyond the range of the heat flux value of interest, the heat flux vectors in the solid shells follow the path along the geometries, and the magnitudes can reach nearly 57 W/m^2 on the side and peripheral of the bottom in the three specimens.

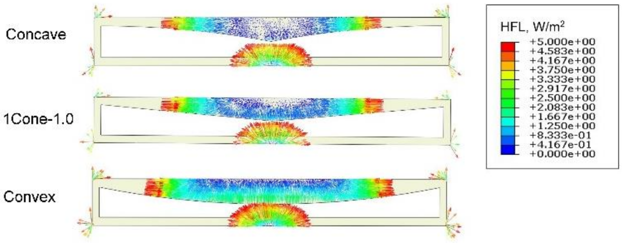


FIGURE 13: HEAT FLUX DISTRIBUTION IN OUTER SHELLS IN THE THREE CONE FEATURED SPECIMENS AT T=2.183 S

In addition, it has been known that the heat flux passing through two contacting objects can be measured as a function of the contact conductance between the two neighboring objects and their temperature gradient. The findings of the heat flux distributed in the shells demonstrate the corresponding tendency of the contact conductance in the three specimens, indicating that the top contact conductance in the Convex specimen exhibits significantly higher than those in 1Cone-1.0 and Concave specimens; although the bottom contact conductance in the three specimens does not vary apparently.

5. CONCLUSIONS

In the present study, hollow specimens were designed with various internal geometries and fabricated with powder enclosed using 17-4PH stainless steel in an L-PBF system. The as-built specimens were tested in a laser-flash equipment to measure the thermal diffusivity up to 500 °C. Then, the combined experimental-numerical simulation was carried out to evaluate the thermal conductivity of the internal powder at 100 °C to 500 °C. An internal geometry effect on the measurement of the metallic powder thermal properties has been analyzed. The contact conditions between the internal powder and the top shell of the specimens exhibited variation between different internal geometrical specimens, which may have resulted in various heat transfer behaviors. The detailed experimental and simulated results are concluded as follows:

- Thermal diffusivity for the 17-4PH specimens with different internal geometries varies from 0.0018 cm^2/s to 0.0025 cm^2/s at 25 °C; wherein, Convex shows the highest values, followed by 1Cone-1.0 and Concave in a descending order at all tested temperatures. As noticed, the thermal diffusivity increases from the room temperature until a plateau at 400 °C, and subsequently increases slightly at 500 °C.
- For each cone configuration specimen, time-response thermograms show an increasing heating rate as temperature rises, with the curves shift to the left, indicative of the rise in sample thermal diffusivity.
- The powder thermal conductivity of 17-4PH is measured ranging from 0.65 W/(m·K) to 1.34 W/(m·K) at 100 °C to 500 °C and displays a linear temperature-dependent tendency for each cone configuration. Specimen geometry has an effect on the powder conductivity evaluation, showing that Convex gives the highest values, and Concave, 1Cone-0.5 and 2Cone-0.25 exhibit similarly lowest.
- As a function of the cone volume, thermal conductivity ratio shows increase with cone volume, although a compromise occurs at the largest cone volume (Convex), and thus appears a non-linear correlation.
- 17-4PH is represented a conductivity ratio of 5.0 % to 5.7 % for 2Cone-0.5 and 4.0 % to 5.0 % for 2Cone-0.25, respectively, which are comparable with Ti64 and IN625 powders at the tested temperatures.
- The bottom contact conductance exhibits similar for the same bottom geometry in different specimens. However, there is variation of the top contact conductance upon

different cone geometries, for example, Convex shows significantly higher top contact conductance. Additionally, the retrogression for both contact conductance in 17-4PH specimens was noticed at 400 °C while this is not the case for Ti64 and IN625.

- Heat flux vectors in the powder of the Convex specimen at t = 2.183 s occurred in a larger region than those in 1Cone-1.0 and Concave specimens. The maximum heat flux value is approximately 5 W/m^2.
- Heat flux in the three cone featured specimens demonstrated the measurement of contact conductance between the internal powder and the outer shell from the developed experimental-numerical approach. Higher top contact conductance in Convex was evaluated than Concave and 1Cone-1.0 specimens, while similar bottom contact conductance was found in the three specimens.

ACKNOWLEDGEMENTS

This research was partially supported by NIST (CA No. 70NANB16H029). The technical support from the computational resource from Cardinal Research Cluster (CRC) at University of Louisville was acknowledged.

REFERENCES

[1] Zhang, Shanshan, Brandon Lane, Justin Whiting, and Kevin Chou, *On thermal properties of metallic powder in laser powder bed fusion additive manufacturing.* Journal of Manufacturing Processes, 2019. **47**: p. 382-392.

[2] *ISO / ASTM52910-18, Additive manufacturing - Design - Requirement, guidelines and recommendations.* ASTM International, West Conshohocken, PA, 2018.

[3] Li, Yali and Dongdong Gu, *Thermal behavior during selective laser melting of commercially pure titanium powder: Numerical simulation and experimental study.* Additive Manufacturing, 2014. **1**: p. 99-109.

[4] Das, Mitun, Vamsi Krishna Balla, Debabrata Basu, Susmita Bose, and Amit Bandyopadhyay, *Laser processing of SiC-particle-reinforced coating on titanium.* Scripta Materialia, 2010. **63**(4): p. 438-441.

[5] Fischer, P, Valerio Romano, Hans-Peter Weber, NP Karapatis, Eric Boillat, and Rémy Glardon, *Sintering of commercially pure titanium powder with a Nd: YAG laser source.* Acta Materialia, 2003. **51**(6): p. 1651-1662.

[6] Dilip, JJS, Shanshan Zhang, Chong Teng, Kai Zeng, Chris Robinson, Deepankar Pal, and Brent Stucker, *Influence of processing parameters on the evolution of melt pool, porosity, and microstructures in Ti-6Al-4V alloy parts fabricated by selective laser melting.* Progress in Additive Manufacturing, 2017: p. 1-11.

[7] Rafi, HK, NV Karthik, Haijun Gong, Thomas L Starr, and Brent E Stucker, *Microstructures and mechanical properties of Ti6Al4V parts fabricated by selective laser melting and electron beam melting.* Journal of materials engineering and performance, 2013. **22**(12): p. 3872-3883.

[8] Zhang, Shanshan, Santosh Rauniyar, Subin Shrestha, Aaron Ward, and Kevin Chou, *An experimental study of tensile property variability in selective laser melting.* Journal of Manufacturing Processes, 2019.

[9] Cheng, Bo and Y Kevin Chou. *Overhang support structure design for electron beam additive manufacturing*. in *ASME 2017 12th International Manufacturing Science and Engineering Conference collocated with the JSME/ASME 2017 6th International Conference on Materials and Processing*. 2017. American Society of Mechanical Engineers Digital Collection.

[10] King, Wayne, Andrew T Anderson, Robert M Ferencz, Neil E Hodge, Chandrika Kamath, and Saad A Khairallah, *Overview of modelling and simulation of metal powder bed fusion process at Lawrence Livermore National Laboratory.* Materials Science and Technology, 2015. **31**(8): p. 957-968.

[11] Whitaker, Stephen, *Local thermal equilibrium: an application to packed bed catalytic reactor design.* Chemical Engineering Science, 1986. **41**(8): p. 2029-2039.

[12] Li, Chi-Hsiung and Bruce A Finlayson, *Heat transfer in packed beds—a reevaluation.* Chem. Eng. Sci, 1977. **32**(9): p. 1055-1066.

[13] Whitaker, S, *Heat and mass transfer in granular porous media.* Adv. Drying, 1980. **1**: p. 23-61.

[14] Klinkenberg, Adrian, *Heat transfer in cross-flow heat exchangers and packed beds.* Industrial & Engineering Chemistry, 1954. **46**(11): p. 2285-2289.

[15] Willhite, GP, Daizo Kunii, and JM Smith, *Heat transfer in beds of fine particles (heat transfer perpendicular to flow).* AIChE Journal, 1962. **8**(3): p. 340-345.

[16] Schotte, William, *Thermal conductivity of packed beds.* AIChE Journal, 1960. **6**(1): p. 63-67.

[17] Masamune, Shinobu and JM Smith, *Thermal conductivity of beds of spherical particles.* Industrial & Engineering Chemistry Fundamentals, 1963. **2**(2): p. 136-143.

[18] Yagi, Sakae and Daizo Kunii, *Studies on effective thermal conductivities in packed beds.* AIChE Journal, 1957. **3**(3): p. 373-381.

[19] Zhou, Jianhua, Aibing Yu, and Yuwen Zhang, *A boundary element method for evaluation of the effective thermal conductivity of packed beds.* Journal of Heat Transfer, 2007. **129**(3): p. 363-371.

[20] Wakao, N and D Vortmeyer, *Pressure dependency of effective thermal conductivity of packed beds.* Chemical Engineering Science, 1971. **26**(10): p. 1753-1765.

[21] Hadley, G R_, *Thermal conductivity of packed metal powders.* International Journal of Heat and Mass Transfer, 1986. **29**(6): p. 909-920.

[22] Vargas, Watson L and Joseph J McCarthy, *Heat conduction in granular materials.* AIChE Journal, 2001. **47**(5): p. 1052-1059.

[23] Wei, Lien Chin, Lili E Ehrlich, Matthew J Powell-Palm, Colt Montgomery, Jack Beuth, and Jonathan A Malen, *Thermal conductivity of metal powders for powder bed additive manufacturing.* Additive Manufacturing, 2018. **21**: p. 201-208.

[24] Bala, Kanan, Pradeep R Pradhan, NS Saxena, and MP Saksena, *Effective thermal conductivity of copper powders.* Journal of Physics D: Applied Physics, 1989. **22**(8): p. 1068.

[25] Cheng, Bo, Brandon Lane, Justin Whiting, and Kevin Chou, *A Combined Experimental-Numerical Method to Evaluate Powder Thermal Properties in Laser Powder Bed Fusion.* Journal of Manufacturing Science and Engineering, 2018. **140**(11): p. 111008.

[26] Zhang, Shanshan, Brandon M Lane, Justin G Whiting, and Kevin Chou. *An Investigation into Metallic Powder Thermal Conductivity in Laser Powder Bed Fusion Additive Manufacturing*. in *Solid Freeform Fabrication Symposium*. 2018.

[27] Clark Iii, LM and RE Taylor, *Radiation loss in the flash method for thermal diffusivity.* Journal of Applied Physics, 1975. **46**(2): p. 714-719.

[28] Ozisik, M Necat, *Inverse heat transfer: fundamentals and applications*. 2000: CRC Press.

[29] Rack, HJ, *Physical and mechanical properties of cast 17-4 PH stainless steel*. 1981, Sandia National Labs., Albuquerque, NM (USA).

[30] *17-4PH stainless steel thermal conductivity. Available from: https://www.upmet.com/sites/default/files/datasheets/17-4-ph.pdf.*

[31] Shurtz, Randy, *Total Hemispherical Emissivity of Metals Applicable to Radiant Heat Testing*. 2018, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

[32] Han, Quanquan, Rossitza Setchi, Sam L Evans, and Chunlei Qiu, *Three-dimensional finite element thermal analysis in selective laser melting of Al-Al2O3 powder.*

[33] Vora, Hitesh D and Narendra B Dahotre, *Multiphysics theoretical evaluation of thermal stresses in laser machined structural alumina.* Lasers in Manufacturing and Materials Processing, 2015. **2**(1): p. 1-23.

[34] Kieruj, Piotr, Damian Przestacki, and Tadeusz Chwalczuk, *Determination of emissivity coefficient of heat-resistant super alloys and cemented carbide.* Archives of Mechanical Technology and Materials, 2016. **36**(1): p. 30-34.

[35] Holman, Jack P, *Heat transfer*. 2010: McGraw-hill.

[36] Cengel, Yunus A, *Introduction to thermodynamics and heat transfer*. Vol. 846. 1997: McGraw-Hill New York.

[37] Rosochowska, M, R Balendra, and K Chodnikiewicz, *Measurements of thermal contact conductance.* Journal of Materials Processing Technology, 2003. **135**(2-3): p. 204-210.

[38] Malinowski, Z, JG Lenard, and ME Davies, *A study of the heat-transfer coefficient as a function of temperature and pressure.* Journal of materials processing technology, 1994. **41**(2): p. 125-142.

[39] Zavaliangos, Antonios, Jing Zhang, Martin Krammer, and Joanna R Groza, *Temperature evolution during field activated sintering.* Materials Science and Engineering: A, 2004. **379**(1-2): p. 218-228.

Compact DC Josephson Voltage Standard

Alain Rüfenacht[*], Anna E. Fox[*], Grace E. Butler[*], Charles J. Burroughs[*], Paul D. Dresselhaus[*],
Robert E. Schwall[*], Stefan Cular[†], and Samuel P. Benz[*]

[*] National Institute of Standards and Technology, Boulder, CO 80305, USA

[†] National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

alain.rufenacht@nist.gov

***Abstract* — This paper presents precision measurements with a prototype cryogen-liquid-free DC Josephson voltage standard that produces a 1 V maximum output. Its cryostat is sufficiently compact that it can be operated on a bench close to devices under test. Compared to the National Institute of Standards and Technology (NIST) programmable Josephson voltage standard (PJVS), this bench-top Josephson voltage standard (BJVS) has a smaller cryostat (only 25 % of the volume), the Josephson junction array circuit is less complex and has only 10 % as many junctions, simpler bias electronics, and reduced mechanical cooling requirements. Direct voltage comparison between the BJVS and the PJVS achieved a relative agreement of 5 × 10^{-10} at 1 V. With the recent redefinition of the International System of Units (SI), the BJVS becomes a primary realization the unit volt, requiring only access to a standard single-phase power outlet and a GPS signal (SI second).**

***Index Terms* — Digital-analog conversion, Josephson arrays, standards, superconducting integrated circuits, voltage measurement.**

I. Introduction

The dissemination of dc voltage from primary to secondary laboratories is nowadays exclusively dependent on the stability of Zener voltage standards. Although the uncertainty achieved with commercial Zener voltage standards is sufficient for most electrical instrument calibration, the intrinsic drift of Zener references requires their periodic calibration with a Josephson voltage standard (JVS) [1]. The Zener calibration with a PJVS is performed at a primary standards laboratory and requires shipping the Zener artifact. With the redefinition of International System of Units (SI), all Josephson voltage standards become a primary realization of the unit volt [2]. In this paper, we present a prototype of a simplified compact bench-top JVS (BJVS) designed to realize the unit volt. The maximum output voltage of the BJVS was chosen to be 1 V, but a small number of fixed voltages are realizable. The main advantages of the 1 V BJVS (Fig. 1) are the reduced complexity and cost of the bias electronics, higher fabrication yield of the BJVS array circuit (having fewer junctions than the 10 V programmable Josephson voltage standard (PJVS)), and the ability to run the cryogenic compressor from a standard single-phase power outlet. This more compact and easily deployable 1 V BJVS can realize the unit volt directly, such as at secondary standards laboratories, eliminating *de facto* multiple calibration measurements, typical of past traceability chains. The BJVS can also be implemented as a primary voltage reference for Kibble balances, which are now key instruments for the realization of the mass scale with the newly redefined SI, or as a traveling primary standard for interlaboratory comparisons.

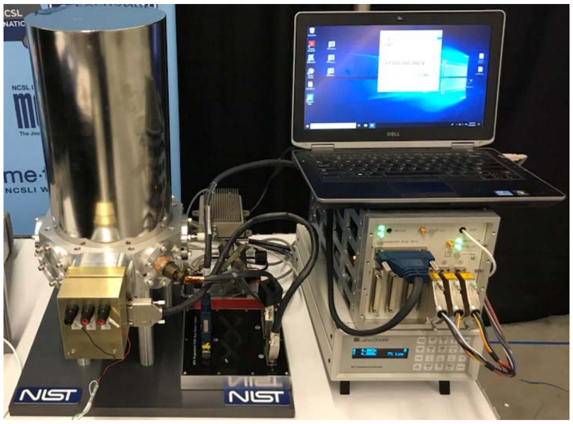


Fig. 1. Cryocooled 1 V bench-top Josephson voltage standard (air-cooled compressor not shown).

II. Bench-Top Josephson Voltage Standard

The BJVS circuit consists of 34 320 Josephson junctions (JJ) biased at a fixed microwave bias frequency of 14.090 8464 GHz. The JJ series array is subdivided into three subarrays of 3 432 JJs, 13 728 JJs, and 17 160 JJs, each generating, when biased on the n=±1 Shapiro step, ±0.1 V, ±0.4 V and ±0.5 V, respectively. The JJ array circuit is mounted on the 4 K cold stage of a Gifford-McMahon cryocooler. The cooling capacity of the cryogenic system is ~150 mW at 4 K, when operated with a single-phase-powered, air-cooled compressor. This cooling capacity is three-times the 50 mW cooling power required to operate the prototype circuit (JJ critical current: $I_c \cong 6.5$ mA at 4 K). A commercial temperature controller coupled to a 50 Ω heater is implemented to maintain the temperature of the device at 4 K. For the bias current sources, we used four channels of a low-cost compact commercial 16-bit digital-to-analog converter (±10 V), as well as four of the 24 current amplifier channels of our PJVS bias electronics amplifier board. A low-cost 24-bit (±0.5 V) analog-to-digital converter module is used to measure the chip output voltage, tune the bias-current operating range, and verify the

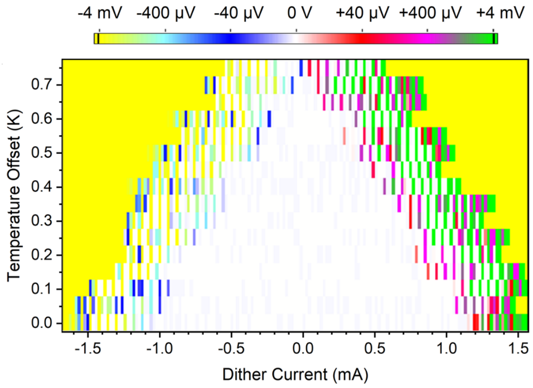


Fig. 2. Quantum locking range for the +0.5 V + (–0.4 V–0.1 V) subarray bias configuration as a function of the dither current and the temperature of the coldhead (referenced from 4 K).

quantum locking range of the BJVS. The microwave frequency is generated by a low-cost compact USB-powered synthesizer having a 100 Hz frequency resolution. The synthesizer is locked to an external 10 MHz reference (NIST primary frequency reference or GPS locked oscillator). The software was adapted to run the BJVS bias electronics while maintaining all the self-checks, diagnostic, and automatic tuning functions previously implemented with the NIST 10 V PJVS.

III. Quantum Locking Range

To measure the quantum locking range (QLR) with the best voltage resolution, the three subarrays are biased in the +0.5 V + (–0.4 V–0.1 V) = 0 V configuration. The white area (Fig. 2) shows the QLR domain as a function of both the dither current and the coldhead temperature. Unlike the 10 V PJVS system [3], the coldhead of BJVS is not fitted with a helium buffer tank, leading to an increase in the temperature oscillations (~240 mK), which are visible at the edges of the QLR by the band of successive voltage variations. Despite the correspondingly small ±0.5 mA bias current margin reduction associated with temperature oscillations, the QLR at 4 K remains larger than 2 mA.

IV. Direct JVS Comparison

We performed a direct comparison at 1 V between the BJVS and the PJVS. A digital nanovoltmeter (DNVM) measured the residual voltage difference. The low and high sides of the DNVM were connected respectively to the low side of the BJVS and the low side of the PJVS. The high sides of both JVS systems were connected in series. The low voltage tap of the BJVS was grounded on-chip through the current bias lead. In this configuration, only the leakage current contribution of the PJVS [4] contributed to a voltage error, but with an overall PJVS leakage resistance >150 GΩ [5], the leakage current-to-ground error contribution to this voltage comparison is insignificant. The measured leakage resistance of the BJVS was 90 GΩ. A single dataset measurement of $\Delta V = V_{\mathrm{BJVS}} - V_{\mathrm{PJVS}}$ consisted of four groups of 15 DNVM readings, where the polarity of both arrays was reversed after every group to remove the thermal offset. All the comparison measurements were performed by applying a dither current sequence (0, +0.2, –0.2) mA in both JVS systems to verify the QLRs [4]. We measured ΔV with two different types of microwave frequency synthesizers on the BJVS system (Table I). If we assume that all the voltage error ($\Delta V = -0.51$ nV) is due to the finite frequency resolution of the compact synthesizer, then its frequency error is ~7 Hz at 14.090 8464 GHz, despite being locked to the same 10 MHz reference as the PJVS synthesizer. With the standard synthesizer replacing the compact synthesizer in the BJVS, this error does not appear and ΔV becomes comparable to the minimum voltage that a typical DNVM can resolve.

Table I: Results of the BJVS-PJVS comparison at 1 V

Synthesizer Type	Synthesizer Resolution	Dataset size	Mean of ΔV (nV)	Type A Unc. k=1 (nV)
Compact	100 Hz	524	**–0.51**	0.37
Standard	1 mHz	357	**–0.14**	0.37

V. Conclusion

The 1 V BJVS is designed as a compact, cost-effective primary voltage standard, for use anywhere with access to a GPS signal needed to establish the traceability of the SI second. With lower complexity and cost compared to the PJVS systems, this BJVS may become a key instrument to disseminate DC voltage at secondary laboratories, shrinking or eliminating the existing traceability chain while improving measurement uncertainty (relative agreement of 5 × 10^{-10} at 1 V). Despite its reduced voltage output, programmability, and tunability (due to fixed frequency), the BJVS ability to generate 1 V and 100 mV enables calibration of a 10:1 voltage divider, which can be combined with the BJVS to extend the input range to 10 V and thereby also enable calibration of 10 V Zener outputs.

References

[1] C. A. Hamilton and L. W. Tarr, “Projecting Zener dc reference performance between calibrations,” IEEE Trans. Instrum. Meas., vol. 52, no. 2, pp. 454–456, Apr. 2003.

[2] A. Rüfenacht, N. E. Flowers-Jacobs, and S. P. Benz, “Impact of the latest generation of Josephson voltage standards in ac and dc electric metrology,” Metrologia, vol. 55, no. 5. pp. S152–S173, Oct. 2018.

[3] A. Rüfenacht et al., “Cryocooled 10 V programmable Josephson voltage standard,” IEEE Trans. Instrum. Meas., vol. 64, no. 6, pp. 1477–1482, June 2015.

[4] A. Rüfenacht et al., “Automated direct comparison of two cryocooled 10 volt programmable Josephson voltage standards,” Metrologia, vol. 55, no. 4, pp. 585–596, Aug. 2018.

[5] C. Burroughs et al., “Automated Leakage Current Measurement Capability for Programmable Josephson Voltage Standards,” Submitted to Proc. Conf. Precision Electromagn. Meas., 2020.

Integrating Field Measurements into a Model-Based Simulator for Industrial Communication Networks

Jing Geng*, Honglei Li*, Mohamed Kashef†, Yongkang Liu†,
Richard Candell†, Shuvra S. Bhattacharyya*
* Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
† Intelligent Systems Division, National Institute of Standards and Technology, USA
Email: {jgeng, honglei}@umd.edu, {yongkang.liu, mohamed.hany, richard.candell}@nist.gov, ssb@umd.edu

***Abstract*— Efficient and accurate simulation methods are of increasing importance in the design and evaluation of factory communication systems. Model-based simulation methods are based on formal models that govern the interactions between components and subsystems in the systems that are being simulated. The formal models facilitate systematic integration across the system, and enable powerful methods for analysis and optimization of system performance. However conventional simulation approaches utilize communication channel models that do not fully reflect the characteristics and diversity of industrial communication channels. To help bridge this gap, we develop in this paper new methods for channel model construction for link-layer simulation that systematically incorporate field measurements of wireless communication channels from industrial networks, and derive corresponding channel modeling library components. The generated library components capture channel characteristics in the form of lookup tables, which can be flexibly integrated into system-level simulators or co-simulation tools. We integrate our new table-generation methods into a model-based co-simulator that jointly simulates the interactions among process flows, physical layouts of workcells, and communication channels in factory systems that are integrated with wireless networks. Experimental results using our lookup-table-augmented co-simulator demonstrate the utility of the proposed methods for flexibly and accurately integrating realistic industrial network channel conditions into simulation processes.**

I. Introduction

The integration of wireless communication capabilities into factory systems is of increasing interest due to the important potential advantages brought about by wireless communications technology in factory environments [1]. This integration leads to highly complex design spaces, which involve interactions among process flow algorithms, factory workcell layouts, and wireless communication networks. We refer to these design spaces as *wireless-integrated factory system* (*WIFS*) design spaces. Due the high complexity of WIFS design spaces, simulation methods are of great importance for designing new systems, and for investigating modifications, such as upgrades, to existing systems. Effective simulation tools enable rapid evaluation and comparison of alternative system designs, thereby facilitating the process of iterative design and system performance optimization.

A limitation in conventional simulation approaches used in navigating WIFS design spaces is that they are based on communication channel models that do not fully reflect the characteristics and diversity of industrial communication channels. Many network simulation tools have been developed whose capabilities are useful to aid in exploring WIFS design spaces. Popular examples include NS-3, OMNeT++, and Cloonix, which provide sophisticated capabilities for network simulation. However, these simulators apply channel abstractions at the physical layer — e.g., by utilizing mathematical equations associated with different channel models or synthetic (simulated) data that is derived from delay profiles obtained from third-party sources, such as the IEEE 802.11 Wireless LAN Working Group (see Section II. WIFS simulation approaches that employ such methods may fail to provide accurate assessment of system-level performance because they do not precisely incorporate characteristics of actual industrial wireless communication channels — for example, harsh conditions that arise due to the vibration of machinery, and the presence of metal objects and obstacles (e.g., see [2]).

In this paper, we develop new methods for integrating realistic models of wireless factory communication network channels into simulation tools and associated WIFS design space exploration processes. More specifically, we introduce an approach to link layer simulation that systematically incorporates data from field measurements. Our link layer simulation approach produces signal-to-noise ratio (SNR) to packet-error-rate (PER) conversion tables, which can be integrated into co-simulation tools for WIFS design space exploration. The resulting integration allows the co-simulation tools to assess factory system performance characteristics (e.g., real-time communication performance, communication reliability, factory production throughput, and workcell energy consumption) more accurately in the context of the reference networks from which the field measurements were taken.

To demonstrate our proposed new approach to integrating field measurements into WIFS design space exploration, we employ a recently-introduced co-simulation tool called *Tau Lide Factory Sim* (*TLFS*), which enables model-based representation and simulation of factory process-flows and systematic integration of its process-flow simulation models with arbitrary discrete event tools for network simulation [3]. We extend TLFS with an extensible channel library that can be populated with different SNR-to-PER lookup tables, such as those generated by our proposed link layer simulation

Geng, Jing; Li, Honglei; Hany, Mohamed; Liu, Yongkang; Candell, Rick; Bhattacharyya, Shuvra. "Integrating Field Measurements into a Model-Based Simulator for Industrial Communication Networks." Paper presented at 16th IEEE International Conference on Factory Communication Systems, Porto, PT. April 27, 2020 - April 29, 2020.

approach. Through experiments, we demonstrate the utility of plugging in SNR-to-PER conversion tables generated from our new link layer simulation tool into TLFS through the new channel library extension. This utility includes both simulation results that more accurately incorporate actual industrial network characteristics, and faster simulation speed, which is enabled by the use of lookup tables as opposed to computational methods.

Although we demonstrate the contributions in this paper using the TLFS co-simulation tools, the utility of the proposed methods is not specific to TLFS. The proposed link layer simulator and the SNR-to-PER tables that it produces can readily be integrated into other network simulators or process-flow / network co-simulation frameworks to extend those tools with more accurate and customizable models for industrial wireless networks.

The remainder of this paper is organized as follows. Section II discusses related work on modeling and simulation of industrial communication networks, and summarizes the contribution of this paper in the context of the related work. Section III presents our new methods for field measurement integration, link layer simulation, and channel characterization library construction. Section III also presents our approach for integrating channel characterization libraries into the TLFS simulation tool, which we use to validate and demonstrate the methods of this paper. Section IV presents experiments involving the results of applying our proposed link layer simulation methods, and results of applying TLFS extended with channel characterizations that are generated using results from these simulation methods. Finally, Section V summarizes the contributions of the paper.

II. Related Work

A large body of work in the literature covers modeling and simulation for cyber-physical systems that are integrated with wireless communication capabilities. These works include two major directions — one that focuses on communication channel modeling for different types of channels, and another that focuses on link layer or system-level simulation that utilizes new channel models or off-the-shelf models, such as those available in MATLAB.

Many approaches have been proposed to model wireless communication channels involving different path loss models, small scale fading models, noise figures, etc. For example, the IEEE 802.11 Wireless LAN Working Group proposes channel models based on a set of WLAN channel prototypes, which model the signal delay profiles for different environments under different indoor/open space line-of-sight (LOS) and non-LOS (NLOS) scenarios (e.g., see [4]). A related class of models is the set of SISO models developed by Medbo and Schramm [5]. However, channel models such as these do not precisely reflect signal transmission in industrial wireless environments. For example, they do not capture harsh characteristics of industrial environments, such as those due to complex surroundings involving vibrating machines and metal structures, which in turn lead to significant multi-path effects, electromagnetic resonance, and other complicating factors (e.g., see [2]).

A variety of works has also investigated channel modeling under more complex communication environments. For example, Abbas et al. present an evaluation of vehicle-to-vehicle communication channel parameters through a detailed comparison between simulations and measurements [6]. Peil et al. use measurements to develop wireless propagation characteristics in an industrial environment, and then use these characteristics to derive a channel model [7].

Other works on modeling and simulation emphasize system-level evaluation based on novel applications of existing simulation frameworks. For example, Liu et al. apply the OMNeT++ simulation library to develop an integrated framework for factory process control simulation and wireless network simulation [8]. Patidar et al. [9] apply link layer simulation to improve the physical layer abstraction. Their approach uses channel profiles proposed from the IEEE WLAN Working Group [4]. Li et al. present a WIFS-oriented design space exploration tool that utilizes TLFS together with channel models that are available in the NS-3 simulation platform [10].

The key distinguishing aspect of our contribution in this paper, compared to the body of literature summarized above, is its focus on coupling the derivation of accurate channel characterizations for industrial wireless environments with system-level co-simulation processes for WIFS design space exploration. This is achieved through a new link layer simulation approach that utilizes field measurements from actual industrial wireless channels, and encapsulation of the resulting channel characterizations through a channel library, which is designed to plug into higher-level co-simulation processes.

III. Approach

In this section, we present our approach for constructing channel models that incorporate the characteristics of industrial wireless networks through field measurements, and applying the derived channel models to enable more realistic co-simulation between factory process flows and wireless communication networks in networked, smart factory environments. We refer to this type of co-simulation as *networked process flow* simulation. The field measurements used in our approach capture channel impulse responses (CIRs) from actual industrial wireless communication environments.

We demonstrate our approach through extensive experiments in Section IV. The field data used in these experiments is publicly available, and was collected from a measurement campaign performed by the U.S. National Institute of Standards and Technology (NIST) [11], [12]. Specifically, we use field data collected in this campaign from an automotive factory site. The field data employed in our experiments is in the form of one large MAT file (approximately 5 GB), which contains all of the measured CIRs from the automotive factory. The MAT file includes the physical location associated with each measured CIR, as well as the IQ data and duration for each CIR signal. The MAT file also includes metadata associated with the measurements, such as the file name, antenna information, and frequency.

In our approach, the measured CIRs are input to a link layer simulator, which in turn produces a PER/SNR table that can be used to efficiently and accurately characterize the communication environment of interest for a higher-level simulator, such as a networked process flow simulator. The approach is extensible so that different wireless environments can be characterized using different PER/SNR tables, which are collected in the form of a channel library, as mentioned in Section I. A channel library populated with tables derived from multiple wireless environments enables comparison of a given factory/network configuration across different environments, as well as comparison of alternative configurations for a given environment.

To model a communication channel, we need to consider factors that include the channel gain, fading, multipath characteristics, and noise, as illustrated in Fig. 1. In our approach, these factors are taken into account by deriving channel models based on the CIR measurement data that is collected from the given factory environment.

Fig. 2 illustrates our link layer simulator, which processes field measurements in the form of CIRs, and produces PER/SNR tables. Our simulator takes as input a set of M location-specific channel impulse responses (CIRs) $C_1, C_2, \ldots, C_M$. Each C_i is associated with a distinct physical location p_i in the factory environment from which measurements are taken. Each CIR C_i is input in the form of an array A_i such that for each sample index j, $A_i[j]$ gives the value of the jth sample in the corresponding CIR signal. The time associated with each of these signal samples can easily be derived by dividing the sample index by the sample rate of the measured signal.

The simulator involves two phases of operation: *model construction* and *simulation*. The horizontal dashed line in Fig. 2 shows the separation of these phases. In the remainder of this section, we describe the different blocks shown in Fig. 2, as well as the operation of the simulation phase.

A. Pre-processing

The Pre-processing block transforms the set of location-specific CIRs $S_c = \{C_1, C_2, \ldots, C_M\}$ into a more compact, *refined* set of location-specific CIRs $S_r = \{R_1, R_2, \ldots, R_N\}$ that are more useful for further analysis than the original set S_c. Here $N \ll M$ since in typical measurement scenarios, M may be very large, and a much smaller value of N is needed for efficient channel library generation as well as for efficient networked factory simulation by tools that use the channel library. To compress across the set of CIRs, we select 1 out every D successive locations that are visited in the measurement process, where D is a parameter of the Pre-processing block. In our experiments, we use $D = 60$.

We also apply pre-processing operations to each individual CIR. In particular, to each CIR, we apply operations for filtering out noise; compression (intra-CIR compression) to reduce the number of samples; and deriving parameters, such as the delay spread and Rician K factor, that compactly characterize each of the CIRs. We apply intra-CIR compression to reduce the computational and memory cost of applying the CIR data. For intra-CIR compression, we first determine the first signal sample whose value exceeds a predetermined threshold, and then extract this sample together with the following 127 samples. The thresholding operation here is performed to ensure that we discard any prefix in the signal that falls within the noise floor, which is assumed not to be part of the desired impulse response. We retain the extracted 128 samples to represent the CIR more compactly and discard all of the other samples. In this context, 128 can be viewed as a particular setting that we use for the *block size* parameter associated with intra-CIR compression in our link layer simulator.

In general, appropriate settings for pre-processing parameters, such as the inter-CIR compression factor and block size parameter, are heavily dependent on the field measurement process. Systematic methods for setting pre-processing parameters based on measurement-process parameters is an interesting direction for future work.

B. Clustering

The Clustering block in Fig. 2 applies the clustering algorithm developed by Kashef et al. [13] to partition the set

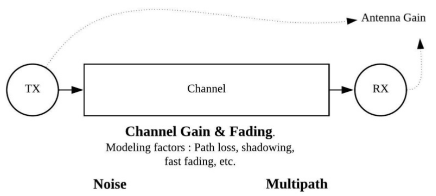


Fig. 1. An illustration of factors needed to model a communication channel.

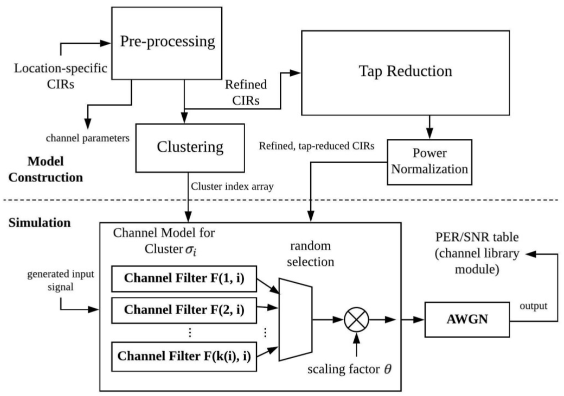


Fig. 2. An illustration of our proposed approach to link layer simulation and channel library generation.

of refined of CIRs into subsets $\sigma_1, \sigma_2, \ldots, \sigma_p$, where all of the CIRs within a given subset σ_i have similar characteristics. Each of the subsets σ_i is referred to as a *cluster* of refined CIRs. After the clustering process is applied, each refined CIR R_i has associated with it a unique cluster, which we denote as $\Sigma(R_i)$. Each cluster σ_i can be expressed as a set $\sigma_i = \{s_{i,1}, s_{i,2}, \ldots, s_{i,k(i)}\}$ of refined CIRs, where $k(i)$ represents the cardinality of (number of elements in) $\sigma(i)$, and $\Sigma(s_{i,j}) = \sigma_i$ for $j = 1, 2, \ldots, k(i)$.

The clustering block also creates a simple data structure, called the *cluster index array*, which maps indices of refined CIRs into indices of the clusters that contain them. Clustering helps to greatly reduce the complexity (number of generated PER/SNR tables) of the output of the link layer simulator, and the corresponding input that is operated on by the networked process flow simulators that utilize this data. This reduction in complexity translates into improved simulation speed during networked process flow simulation, as well as a more streamlined process for creating and storing the channel library. The reduction in complexity is achieved while preserving key characteristics of the original (unclustered) set of refined CIRs based on properties of Kashef's clustering algorithm.

C. Tap Reduction and Power Normalization

The tap reduction block converts each refined CIR R_i into a more compact tapped delay profile T_i using the algorithm of Mehlfuhrer and Rupp [14]. This is done, as with the inter-CIR and intra-CIR compression processes described above, to reduce model complexity in a way that does not substantially diminish the utility of the model. The number of taps in each of the T_is is a parameter of the tap reduction block that controls the complexity/accuracy trade-off of the tap-reduced CIR form. We refer to this as the *reduced tap count parameter* of our link layer simulator. In our experiments, we use 18 as the value of this parameter.

The output of the tap reduction process is post-processed by a simple power normalization block, as shown in Fig. 2. This block converts the taps to a form in which the sum of their squares equals unity. This normalization is performed to provide a channel profile without any path loss.

D. Construction of Cluster-Level Channel Models

Each cluster σ_i derived by the clustering block is converted during the model construction phase into a channel model, as illustrated by the block in Fig. 2 that is labeled Channel Model for Cluster σ_i. This block consists of a bank of filters $F(1,i), F(2,i), \ldots, F(k(i),i)$, where, as defined in Section III-B, $k(i)$ is the number of CIRs in cluster σ_i. Each $F(j,i)$ is the tap-reduced, power-normalized form of the refined CIR $s_{j,i}$.

Fig. 3 illustrates an example of a single filter $F(j,i)$ based on data from field measurements, and our reduced tap count parameter setting of 18. Section IV provides more details about the field measurement data that we have employed in our examples and experiments. The simulation model consists of $(k(1) + k(2) + \ldots k(p))$ such filters, where the filters are grouped into their corresponding clusters.

E. Simulation of the Constructed Channel Model

We use MATLAB to prototype our link layer simulator using the simulation model construction approach illustrated in the top part of Fig. 2. As a starting point for the prototype, we use the MATLAB WLAN toolbox, which contains channel modeling capabilities based on models proposed by Erceg et al. [4]. We then make various modifications and extensions to realize the proposed new simulation capabilities for bridging field measurements to networked process flow simulation.

First, we bypass the multipath features and fading profile from the MATLAB simulation. Instead, in our simulation approach, the multipath and fast fading effects are modeled through clusters and filters derived from the field measurements. For large scale fading, we apply the two-slope path-loss model proposed in by Damsaz et al. [15]. This is achieved in the simulation by scaling the filter response with a scaling factor θ whose value varies during simulation based on a randomly determined shadowing parameter. During simulation, applying this scaling factor is equivalent to multiplying the filter output by θ, as illustrated by the multiplication block shown in the lower part of Fig. 2.

In our approach, the link layer simulator is used to create a channel library module associated with a given factory environment. Each channel library module is populated with a set of PER/SNR tables, where one table is produced corresponding to each cluster. Different communication system parameters (e.g., the modulation and coding scheme or MCS) may be varied to produce a parameterized set of tables. In this case, for each set Γ of relevant parameter settings/combinations, link layer simulations are executed to derive the PER/SNR tables (table-subset) associated with Γ. Then the desired table-subset can be selected and used as input during networked process flow simulation.

To generate the table-subset for a given cluster σ_i and a given set of communication system parameters, the channel model for σ_i — based on the associated filters $F(1,i), F(2,i), \ldots, F(k(i),i)$ — is constructed as the core of the simulation model, as shown in the bottom part of Fig. 2. The simulation constructs physical layer service data unit (PSDU) signals from a stream of randomly-generated

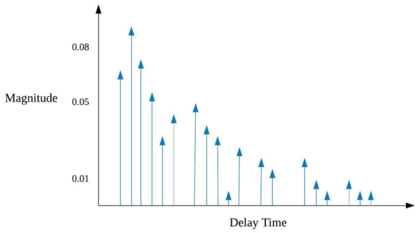


Fig. 3. An example of a single filter based on data from field measurements.

packets. The PSDU signals are passed through the channel model for σ_i to calculate the packet error rate.

For each packet, one of the $k(i)$ channel filters is selected randomly from $\phi_i = \{F(1,i), F(2,i), \ldots, F(k(i),i)\}$. This random selection allows us to incorporate the transient performance of the communication channel into the simulation in a realistic manner. The simulation to determine the packet error rate is run over a large number of packets and for a specific SNR value. This process is repeated over different SNR values to obtain the PER/SNR table for cluster σ_i.

The random selection from the set of filters ϕ_i is represented by the randomly-controlled switch that is connected to the output of the channel filters in Fig. 2. The output of the switch is connected to a multiplication block to simulate large scale fading, as described above.

Fig. 4 illustrates the use of the filter set ϕ_i to model a communication channel based on the corresponding cluster σ_i.

F. Integration with Networked Process Flow Simulation

As described in Section I, we use the TLFS tool for networked process flow simulation to demonstrate the utilization of a channel library produced from our link layer simulation approach. The model-based architecture of TLFS facilitates its extension with novel capabilities, such as those presented in this paper for channel modeling. Fig. 5 illustrates the high-level architecture of a TLFS-based networked process flow simulation together with the new TLFS extension to utilize channel libraries that are provided by external tools, such as our link layer simulator.

Here, the factory process flow is modeled (by the simulation tool user) as a dataflow graph, as illustrated in the bottom left part of Fig. 5. The dataflow graph illustration here incorporates various *actors* (dataflow-based functional components) that are useful for simulating factory process flows in TLFS. The actors labeled M and R represent a machine and a rail, respectively, while the actor labeled C is a controller actor that models the control of a set of machine and rail components. The dataflow graph also contains a number of send interface actors (SIAs) and receive interface actors (RIAs) which provide model-based interfaces between the process flow (dataflow) model and the communication network (discrete event) model.

The model-based architecture of TLFS makes it possible to plug in different network simulators to enable co-simulation between such simulators and the dataflow-based process flow simulator within TLFS. In our experiments, we utilize the popular NS-3 [16] simulator as the network simulator, as illustrated by the lower right part of Fig. 5. The blocks labeled NN and AP in this part of the figure respectively represent models for network nodes and an access point. In the extended version of TLFS illustrated in Fig. 5, we configure the network simulation to bypass the PER modeling features built-in to NS-3 and utilize instead the PER/SNR tables provided by the channel library.

For more details on the TLFS-based modeling and co-simulation methods illustrated in Fig. 5, see [3].

IV. Experiments

In this section, we present experiments that demonstrate the utility of our proposed linked layer simulation approach and its integration with networked process flow simulation.

A. Field Measurement Dataset

For the set of field-measurements that are input our link layer simulation experiments, we apply CIR measurements that have been obtained from a measurement campaign performed by NIST, as mentioned in Section III. Details of the measurement techniques involved in this campaign are presented by Candell et al. [11]. We specifically use the field data collected in this campaign from an automotive factory site. Among the different sites surveyed in the measurement campaign, the automotive factory most closely matches the factory production environment context to which this paper is most oriented. All of the field data is archived online [12].

Applying Kashef's clustering technique [13] on the CIR data from the automotive plant, after pre-processing, yields a total four clusters. Fig. 6 shows the *representative CIR* that is derived from each of the four clusters. The representative

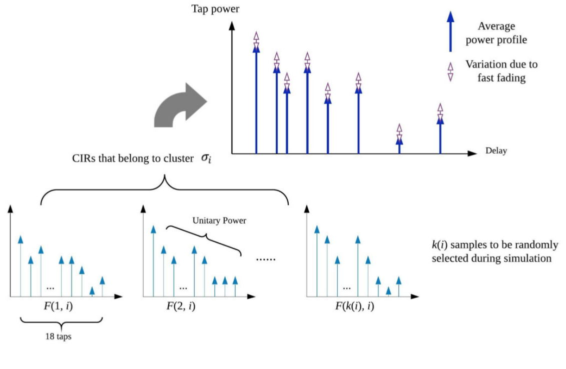

Tap power
Average power profile
Variation due to fast fading
Delay
CIRs that belong to cluster σ_i
Unitary Power
...
......
$k(i)$ samples to be randomly selected during simulation
$F(1, i)$
$F(2, i)$
$F(k(i), i)$
18 taps

Fig. 4. An illustration of the use of the filter set ϕ_i to model a channel based on cluster σ_i.

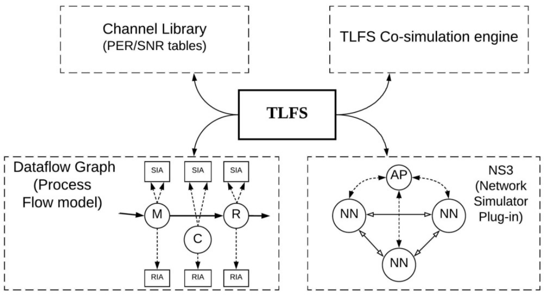


Fig. 5. An illustration of the architecture of a TLFS-based networked process flow simulation together with the new TLFS extension to utilize channel libraries.

CIR of a cluster is simply the CIR that results from averaging all of the CIRs in the cluster. We denote the clusters whose representatives are shown in Fig. 6a–d, respectively, by $\sigma_1, \sigma_2, \sigma_3, \sigma_4$. Cluster σ_1 corresponds to an NLOS channel in the factory, while clusters $\sigma_2, \sigma_3, \sigma_4$ correspond to LOS situations with different delay spreads. Differences in impact on channel performance due to CIRs from different clusters are demonstrated in Section IV-E.

B. *Experimental Setup for Comparing PER/SNR Curves*

In Section IV-C through Section IV-E, we present relevant experimental results and comparisons involving PER/SNR tables that are derived using our proposed linked layer simulation approach. These experiments are performed using the following parameters.

- Duration of simulation. For each SNR point in each generated table, we simulate until there are 10^5 packets transferred through the given channel or until 10^4 packet errors are detected. Whichever condition is detected first triggers the end of the simulation.
- Reduced tap count parameter. The value of this parameter is set to 18 taps, as stated in Section III-C.
- Protocol. All of the packets are generated randomly and transmitted using the 802.11n protocol with 2.4GHz frequency.
- Packet size. A packet size (PSDU length) of 100 bytes is used in all simulations.

C. *Comparison with TGn Channel*

In this subsection, we present a comparison between the channel model produced by our link layer simulator and the TGn model [4], which we choose here as a well-known, representative example of a model that is not specific to industrial environments. The differences between results using these two models help to concretely demonstrate the importance of customizing the channel model for an industrial environment when this type of environment is being studied.

Fig. 7 compares the PER/SNR curves generated from our link layer simulator using the measured automotive plant data with simulation results using the TGn model. To present the results without excessive clutter, we select one representative cluster (σ_3), and three TGn delay profiles (B, D, and E). In our settings for simulation, delay profiles B and D correspond to NLOS scenarios, while E is for LOS. The modulation and coding scheme (MCS) is set to MCS4 for all simulations. For a summary of all of the MCS indices for 802.11n, we refer the reader to [17]. The blue curve shows the PER/SNR characteristic derived by simulating cluster σ_3 in our link layer simulator, while the other three curves show PER/SNR results that we obtained by simulating the selected TGn delay profiles in MATLAB.

From Fig. 7, we can see that there can be significant difference between the performance of a real industrial channel and a TGn channel, and that this difference is captured by our link layer simulator. The disparity illustrated in Fig. 7 helps to quantitatively motivate the need for more accurate integration of channel characteristics into networked process flow simulations and related types of simulations for industrial wireless environments.

D. *Comparison Involving Different MCS Settings*

As mentioned in Section III-E, our link layer simulator can generate a parameterized family of PER/SNR tables based on designer-specified communication parameters. In this section, we demonstrate the utility of this capability by applying the MCS as a parameter for table generation. Fig. 8 shows the generated PER/SNR curves for cluster σ_3 (from the automotive factory) for MCS settings 0 through 6. Here, we use only these 7 MCS settings because the field measurements used in our experiments are based on a SISO antenna. However, our simulation framework of Section III is not restricted to the SISO case. Indeed, it can be applied to MIMO channels. Due to use of field measurements as opposed to mathematical models, the approach in our simulator for handling MIMO channels does not involve traditional MIMO modeling techniques based on correlation matrices (e.g., see [18]).

The results shown in Fig. 8 show that MCS selection can have significant impact on PER. Thus, in the context of WIFS design space exploration, it is useful to have available a parameterized collection of PER/SNR tables where one can easily vary the MCS scheme to assess the overall impact on

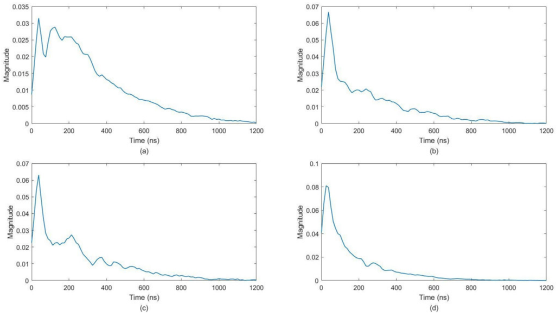


Fig. 6. Representative CIRs derived from the four clusters obtained from the automotive plant measurements. Parts (a)–(d) show the representative CIRs for clusters σ_1 through σ_4, respectively.

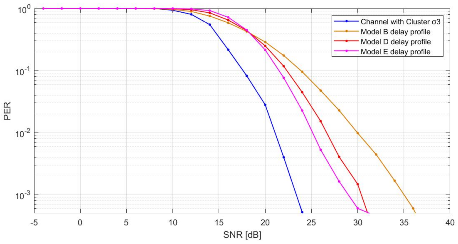


Fig. 7. A comparison between the PER/SNR curves generated from our link layer simulator using the measured automotive plant data with simulation results using the TGn model.

Geng, Jing; Li, Honglei; Hany, Mohamed; Liu, Yongkang; Candell, Rick; Bhattacharyya, Shuvra. "Integrating Field Measurements into a Model-Based Simulator for Industrial Communication Networks." Paper presented at 16th IEEE International Conference on Factory Communication Systems, Porto, PT. April 27, 2020 - April 29, 2020.

factory system performance in terms of PER and other relevant metrics.

E. Comparison Among Different Clusters

Fig. 9 illustrates the different PER/SNR curves generated by our link layer simulator for the four clusters in the automotive factory dataset. The MCS setting used in this experiment is 4. Cluster σ_1 exhibits poor PER due to its NLOS characteristic and large delay spread. Clusters $\sigma_2, \sigma_3, \sigma_4$, on the other hand, correspond to LOS scenarios. Among these, σ_4 has the lowest PER since signals transmitted in this type of channel have high K-factor, and low delay spread quality. Clusters σ_2 and σ_3 show similar performance characteristics in Fig. 9 with Cluster σ_3 exhibiting slightly better performance.

The results in Fig. 9 illustrate how networked factory simulation results may motivate changes to a factory layout. For example, if better communication reliability is desired than what an NLOS channel can support in a given deployment, then a rearrangement of the layout may be performed to provide a "clean" physical environment for the relevant communication path. This may lead to higher quality PER/SNR curves, which, when plugged into to the networked factory simulator, may provide the desired level of estimated reliability.

F. Integration into Networked Process Flow Simulation

In this section, we present experiments involving the integration into a networked process flow simulator of PER/SNR tables generated by our link layer simulator. As mentioned previously, the networked process flow simulator that we use to demonstrate this integration is TLFS. The experiments presented in this section help to validate the implementation of channel libraries for TLFS based on results of the new link layer simulator, and to demonstrate the kinds of networked process flow simulations and WIFS design space exploration that can be carried out based on field measurements, as enabled through the channel libraries.

The factory process flow model used in these experiments is a pipelined structure with one parts generator, one parts sink, three machines, four rails, and three machine controllers. The parts sink actor represents a subsystem that collects and stores parts after they are processed by the pipeline. For details on the modeling of rails, machines, controllers and other types of factory subsystems in TLFS, see [3].

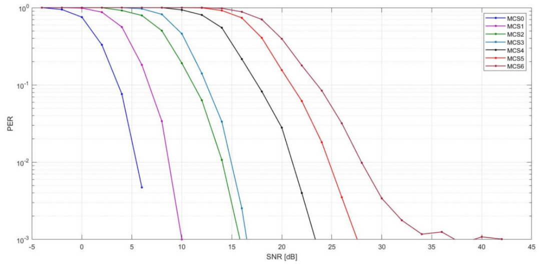


Fig. 8. A comparison across different MCS settings for cluster σ_3.

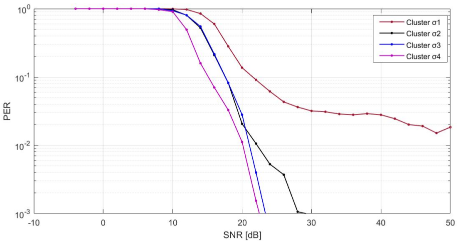


Fig. 9. PER/SNR curves generated by the link layer simulator for the four clusters in the automotive factory dataset.

For simulation, the machine working time (the time for a machine to process a given part) is determined randomly from a uniform distribution within the range $[0.9\mu, 1.1\mu]$, were μ is a model parameter that is set to 40 sec. in our experiments. When initializing the process flow model for each simulation run, the physical spacing between the machines, rails, and machine controller was also determined randomly: the distance between adjacent subsystems was determined from a uniform distribution on $[0.9\delta, 1.1\delta]$, where $\delta = 2 \text{ meters}$. In each simulation run, 10 products were generated by the parts generator actor and processed through the complete pipeline.

Fig. 10 shows networked process flow simulation results for the pipeline model described above using the generated PER/SNR tables associated with different MCS settings. Here we fixed the cluster to be σ_3, and varied MCS indices across the set $\{0, 1, 3, 6\}$. Here, to present the results without too much clutter, we have selected a subset of four representative MCS indices rather than evaluating all of them. For each MCS index, we ran 10 TLFS simulations independently and averaged the results over all the packets. The communication protocol employed in the experiment was the 802.11n protocol.

As shown in Fig. 10, the average communication delay decreases with increasing MCS indices, which is consistent with the increasing data rates associated with higher indices.

Fig. 11 shows networked process flow simulation results using the generated PER/SNR curves for the different clusters $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ under a fixed MCS setting of 4. Here, we have randomly selected 4 as a representative MCS index to show results for a fixed index. In this experiment, we increase the mean spacing parameter to $\delta = 18 \text{ meters}$. At this increased spacing, the SNR falls into a range where there is significant variation among the PER levels across the different clusters. All other parameters are kept the same as in the experiments associated with Fig. 10. The results in Fig. 11 are consistent with those in Fig. 9. We see that The PER/SNR curve for σ_4 leads to the lowest average communication delay and the one with σ_1 gives the highest delay.

V. Conclusions

In this paper, we have developed new methods for integrating realistic models of wireless factory communication network channels into simulation tools and design space exploration processes for wireless-integrated factory systems. Our approach involves systematically incorporating data from field measurements into link layer layer simulation, and generating channel libraries that accurately incorporate channel characteristics into networked process flow simulation. We have demonstrated the proposed methods using a large dataset containing field data collected from an automotive factory. Interesting directions for future work include systematic methods for setting pre-processing parameters based on measurement-process parameters; integration of channel libraries into automated design optimization processes for factory process flows; and application of channel libraries to design space exploration for other types cyber-physical systems.

Disclaimer

Certain commercial equipment, instruments, materials, software or systems are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

[1] A. A. Kumar S., K. Ovsthus, and L. M. Kristensen, “An industrial perspective on wireless sensor networks — a survey of requirements, protocols, and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 3, pp. 1391–1412, 2014.

[2] K. Wiklundh, “Interference challenges for industry communication,” 2019, pDF presentation slides from keynote at WFCS 2019, downloaded from https://www.miun.se/en/thank-you-for-participating on 01/22/2020.

[3] J. Geng *et al.*, “Model-based cosimulation for industrial wireless networks,” in *Proceedings of the IEEE International Workshop on Factory Communication Systems*, 2018, pp. 1–10.

[4] V. Erceg, “TGn channel models,” IEEE P802.11 Wireless LANs, Tech. Rep. IEEE 802.11-03/940r4, 2004.

[5] J. Medbo and P. Schramm, “Channel models for HIPERLAN/2 in different indoor scenarios,” Ericsson Radio Systems AB, Tech. Rep. 3ERI085B, 1998.

[6] T. Abbas, J. Nuckelt, T. Kürner, T. Zemen, C. Mecklenbräuker, and F. Tufvesson, “Simulation and measurement based vehicle-to-vehicle channel characterization: Accuracy and constraint analysis,” 2014, arXiv:1410.4187v1 [cs.NI].

[7] J. Peil, M. Damsaz, D. Guo, W. Stark, R. Candell, and N. Moayeri, “Channel modeling and performance of Zigbee radios in an industrial environment,” National Institute of Standards and Technology, Tech. Rep., September 2016.

[8] Y. Liu, R. Candell, K. Lee, and N. Moayeri, “A simulation framework for industrial wireless networks and process control systems,” in *Proceedings of the IEEE World Conference on Factory Communication Systems*, 2016, pp. 1–11.

[9] R. Patidar, S. Roy, T. R. Henderson, and A. Chandramohan, “Link-to-system mapping for ns-3 Wi-Fi OFDM error models,” in *Proceedings of the Workshop on ns-3*, 2017, pp. 31–38.

[10] H. Li, J. Geng, Y. Liu, M. Kashef, R. Candell, and S. Bhattacharyya, “Design space exploration for wireless-integrated factory automation systems,” in *Proceedings of the IEEE International Workshop on Factory Communication Systems*, May 2019, 8 pages in online proceedings.

[11] R. Candell *et al.*, “Industrial wireless systems radio propagation measurements,” National Institute of Standards and Technology, Tech. Rep. 1951, 2017.

[12] “Networked control systems group — measurement data files,” https://www.nist.gov/el/intelligent-systems-division-73500/networked-control-systems-group/measurement-data-files, 2020, visited in January 2020.

[13] M. Kashef, R. Candell, and Y. Liu, “Clustering and representation of time-varying industrial wireless channel measurements,” in *Proceedings of the Annual Conference of the IEEE Industrial Electronics Society*, 2019, pp. 2823–2829.

[14] C. Mehlfuhrer and M. Rupp, “Approximation and resampling of tapped delay line channel models with guaranteed channel properties,” in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2008, pp. 2869–2872.

[15] M. Damsaz, D. Guo, J. Peil, W. Stark, N. Moayeri, and R. Candell, “Channel modeling and performance of Zigbee radios in an industrial environment,” in *Proceedings of the IEEE World Conference on Factory Communication Systems*, 2017, pp. 1–10.

[16] *ns–3 Tutorial, Release ns–3.25*, ns–3 Project, 2016.

[17] “MCS index for 802.11n and 802.11ac chart,” https://www.wlanpros.com/resources/mcs-index-802-11ac-vht-chart/, 2020, visited in January 2020.

[18] L. Schumacher, K. I. Pedersen, and P. E. Mogensen, “From antenna spacings to theoretical capacities — guidelines for simulating MIMO systems,” in *Proceedings of the International Symposium on Personal, Indoor and Mobile Radio Communications*, 2002, pp. 587–592.

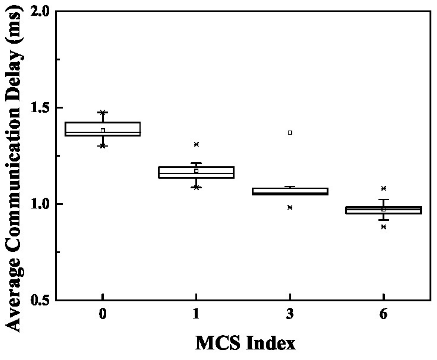


Fig. 10. An example of networked process flow simulation results using PER/SNR data from a channel library produced by our link layer simulator.

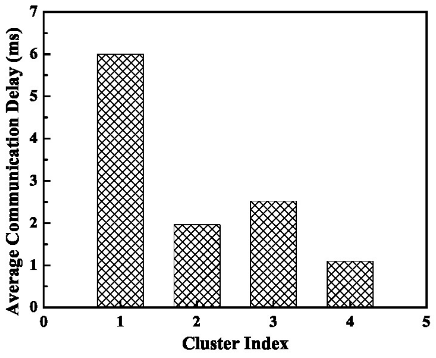


Fig. 11. Networked process flow simulation results for the four different clusters under $\mathrm{MCS} = 4$.

Dynamic Flow Stress Measurements of 6061-T6 Aluminum under Rapid Heating for Machining Studies

Homar Lopez-Hawa[1], Steven P. Mates[2], Wilfredo Moscoso-Kingsley[1], Viswanathan Madhavan[1]
[1]Wichita State University, 1845 Fairmount St, Wichita, KS, USA
[2]National Institute of Standards and Technology, Gaithersburg, MD, USA.

ABSTRACT

The flow stress of aluminum alloy Al6061-T6 produced by conventional thermomechanical methods has been investigated under the application of strains up to 80 % at a strain rate of the order of 10^3 1/s, and temperatures ranging from ambient to near melting. The high strain and strain rate deformation was imposed using a Kolsky bar apparatus equipped with a fast pulse-heating system to reach the target temperatures at high heating rates. The flow stress was measured to provide a constitutive model under strains, strain rates, temperatures and heating rates that match thermomechanical conditions developed in the primary shear zone for specially-designed comparative machining tests. It is expected that these measurements will enable the formulation of realistic machining models for Al6061-T6. Temperature measurements were obtained using non-contact infrared (IR) full field imaging together with embedded micro-thermocouple (TC) point probing, the latter being used to estimate the emissivity of the aluminum sample via separate, *in situ* calibration tests. Type-K TC measurements were made using the separated junction principle by embedding the TC wires into two micromachined holes in the specimen. The temperature measurement technique is discussed in detail, and temperature uncertainties are estimated. The technique will be used going forward to study the effect of heating time on the dynamic thermal softening behavior of Al6061-T6, which has been shown to be time-sensitive under quasi-static loading when temperatures exceed 200 °C due to Mg-Si precipitate growth.

KEY WORDS: Aluminum, Plastic Behavior, Kolsky Compression, Machining, High Strain, High Strain Rate, High Temperature, High Heating Rate.

INTRODUCTION

Manufacturing processes impose extreme strains, strain rates and temperatures [1]. Because these processes are designed to transform materials, it is important to understand material behavior under such extreme conditions. In particular, for machining, accurate information about the mechanical response of the workpiece material is required to obtain predictive models to optimize performance measures such as surface roughness, energy expenditure and tool wear [2]. One well-known experimental approach that provides accurate information about mechanical response (flow stress) under extreme thermomechanical conditions is the Kolsky bar compression technique. With this technique, dynamic compression (high strain and strain rate) that is similar, in terms of peak strains and strain rates, to deformation imposed by machining may be produced. The Kolsky compression tests may be performed under elevated temperature. However, the heating rates commonly employed in this kind of material testing are quite low compared to those occurring during machining. As a result, the flow stress data obtained may reflect equilibrium material conditions that are quite different than the transient conditions present during machining if the material transforms under heating. For example, in aluminum alloy Al6061, Mg-Si precipitates can grow above 200 °C, causing a significant change in flow stress depending on the time-at-temperature [3].

This work presents results from dynamic compression experiments performed on a model, precipitation treatable alloy (aluminum alloy Al6061-T6), with a special Kolsky bar apparatus equipped with an electric pulse heating element. The compression was performed at high strains and strain rates (up to 80 % and 10^3 1/s, respectively), and at elevated temperatures (up to 500 °C) under rapid heating (of the order of 1000 °C/s). These conditions were selected to match strain, strain rate, temperature and heating rate achievable by special machining tests, where the strains were lowered to less than 100 % by the combined application of highly-positive rake angles and chip pulling [4]. The work is aimed at the generation of material models that include contributions from non-equilibrium, time-sensitive, phase

Mates, Steven P.; Lopez-Hawa, Homar; Moscoso-Kingsley, Wilfredo; Madhavan, Viswanathan. "Dynamic Flow Stress Measurements of 6061-T6 Aluminum under Rapid Heating for Machining Studies." Paper presented at Society for Experimental Mechanics 2020 Annual Meeting, Orlando, FL, US. June 08, 2020 - June 11, 2020.

transformations that might be more suitable for simulating machining processes compared to models developed from more conventional high temperature Kolsky bar techniques.

EXPERIMENTAL

An extruded 6061-T6 aluminum block purchased commercially was used as the specimen source. The specimens were cut from this block using electrical discharge machining (EDM). Disk-shaped specimens measuring 4 mm in diameter and 2 mm in thickness were dynamically compressed with a Kolsky bar system under both ambient temperature conditions and with rapid pre-heating using a pulse-direct current heating method [5]. Dynamic compression experiments were conducted with a 375 mm striker launched pneumatically at a gage pressure of 205 kPa (30 psi), which produced plastic strains up to 80 % at strain rates of the order of 10^3 1/s. For smoothing the incident pulse, copper pulse shapers, measuring 6.35 mm diameter and 0.0254 mm thick (¼ inch diameter and 0.01 inch thick), were placed in between the striker bar and the incident bar. The strain pulses were measured by metal foil strain gages connected to a battery-powered bridge circuit. The output was sampled at 2 MHz using a digital oscilloscope.

For carrying out the heated Kolsky bar experiments, a single short direct current pulse, produced by a low voltage, high current capacity battery bank (12 V), was passed through the specimen and the bars. Samples were heated using controlled current pulses of 35 A to 90 A with the times ranging from several seconds down to 0.2 s in order to obtain a range of temperatures and heating times prior to impact. With 35 A being the lowest controllable current for the equipment and specimen size used in this study, the minimum achievable steady-state temperature was about 360 °C. Since precipitate growth becomes an issue at temperatures as low as 200 °C, it was of interest to explore lower temperatures, which were achieved by short, ramped current pulses with durations less than 1 second. Because of the very high thermal conductivity of aluminum, the temperature uniformity of the specimens was good even during these very short ramped experiments.

Temperature measurements were obtained using an *in situ* calibrated mid-wave infrared (MWIR) thermal camera, with operating wavelengths between 1.5 µm to 5 µm, a sensor integration time of 150 µs, a frame rate of 870 Hz, and a sensor resolution of 160 pixels by 128 pixels with a magnification yielding 38 µm/pixel. The camera was first calibrated using a blackbody furnace, followed by a second *in situ* calibration that consisted of imaging heated aluminum samples while simultaneously measuring the temperature with a fine-wire (0.127 mm, 0.005 in diameter) type-k thermocouple embedded within the sample. Embedding the thermocouple was necessary because the individual wires were too fine to be reliably spot-welded onto the surface. Embedding the wires involved carefully drilling two small holes into the side wall of the sample using a #80 drill mounted on a linear micrometer stage (Figure 1, left). Each thermocouple wire was press-fit into a hole using a second, small piece of the same thermocouple wire as a wedge. An installed thermocouple can be seen in Figure 1, right. With this distributed-junction thermocouple technique, sample temperature measurements were obtained during the *in situ* thermal camera calibration tests to determine the effective surface emittivity (emissivity plus scattering) for the aluminum samples in the Kolsky bar measurement environment. While it was also determined that the thermocouple holes did not significantly affect the flow stress measurement by comparing identical compression tests on samples with and without holes, it was decided to use the calibrated thermal camera to determine impact temperatures for heated compression tests to avoid having to drill each sample.

The stress-strain curves were analyzed using the usual wave analysis methods [5] with two exceptions. First, because graphite foil was used as a lubricant in the heated tests as well to avoid electrical arcing during heating, the mechanical contribution of the foil was subtracted from the total mechanical response to obtain the specimen response. The complete method of graphite foil correction can be found in the work done by Mates et al. [5]. The other correction implemented was the correction of the elastic indentation of the bars during the dynamic compression, using the method described by Safa and Gary [6]. However, this correction is almost negligible for aluminum.

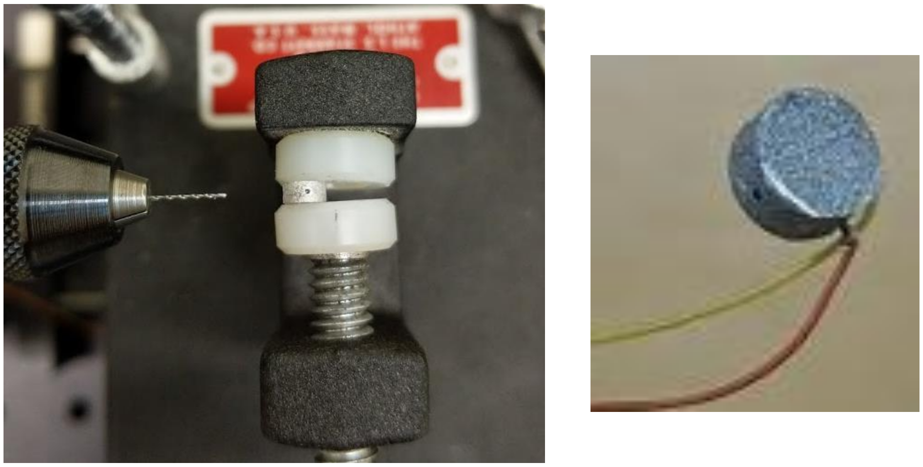


Figure 1. Procedure for drilling holes for the micro-thermocouple point probing installation on a 2 mm thick by 4 mm diameter aluminum sample.

RESULTS

Figure 2 illustrates the flow stress of Al6061-T6 at essentially ambient temperature, without any current going through the sample prior to or during the test. Any temperature rise during this test is only from dissipation of plastic work arising from the compression. Estimates based on temperature measurements using the embedded thermocouple indicate this temperature rise to be no more than 50 °C. In these experiments, grease lubricant was used instead of the foil, and therefore the stress-strain analyses were performed with only the elastic indentation correction applied. The drop-off in the flow stress for strains close to 0.8 is caused by the release-wave, which is produced at the end of the experiment, and does not represent material failure. Also, the upturn in the curve starting from strain about 0.7 may be attributed to friction. The curve shown in the figure represents the mean flow stress of 9 experiments performed under identical impact conditions. The error bars correspond to three times the standard deviation of the flow stress at each strain point. The flow stress observed corresponds well to previous studies of aluminum alloys [7]. The values measured also correspond well with flow stress estimates from special machining tests performed with the intention of matching the strain and strain rate of our Kolsky bar tests performed at ambient temperature [4]. These machining tests were designed to match strains and strain rates in the Kolsky bar tests. The machining tests performed at strain of 0.6, strain rate 10^2 1/s and essentially ambient temperature resulted in flow stress of 383 MPa [4]. This value is within the measurement error in flow stress shown in Figure 2. The agreement between the machining test and the Kolsky bar data is good despite the differences in stress states between the two types of test. The strain rate produced by the machining test was an order of magnitude lower than that produced by the Kolsky bar test. Future machining tests will be performed at faster cutting speeds to improve the strain rate match. However, flow stress estimates are not expected to change significantly, as it is known that the strain rate sensitivity of Al6061-T6 is low at strain rates less than 10^4 1/s [7].

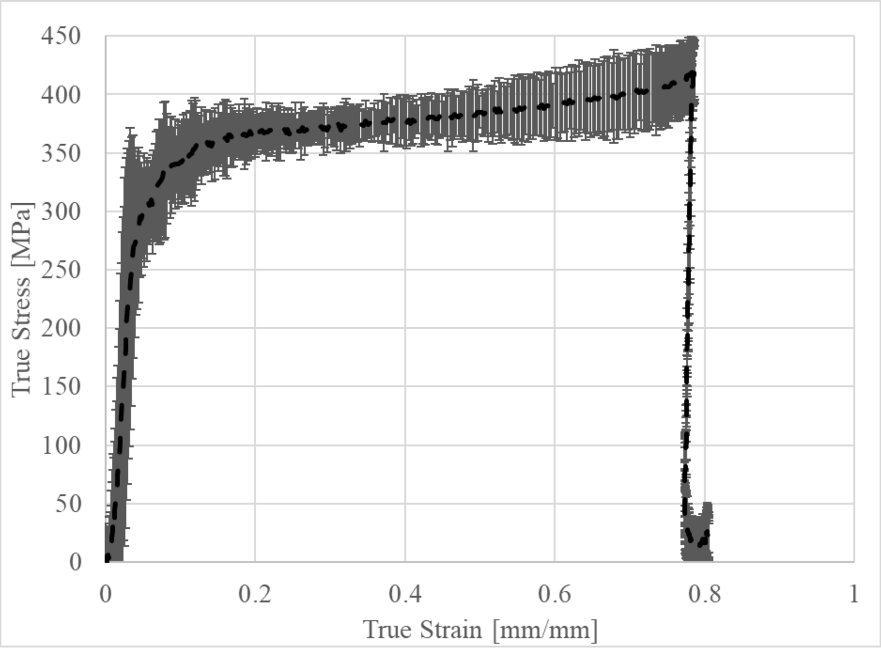


Figure 2. Flow stress as a function of strain. The flow stress is estimated by an average based on 9 repeated experiments. The error bars correspond to ± 3 standard deviations.

Figure 3 shows the plastic flow stress at several temperatures up to about 500 °C at a fixed level of plastic strain (50 %). The heating was performed by applying current pulses of different dwell times (heating times). However, the heating time was always very short – with a maximum of 7.5 s. The higher the temperature and the longer the heating time, the smaller the flow stress (Figure 3). This result is generally reflective of what has been shown at lower strain rates, where time-sensitive thermal softening is likely caused by precipitate growth [3], although grain growth may also play a role as well. In reference [3], a significant reduction in flow stress was reported when a test was performed after subjecting the material to a thermal treatment. The thermal treatment had for target temperature 400 °C. The heating rate was 200 °C/s, and the cooling rate was of the order of 10 °C/s. Therefore, the total time under temperature above 275 °C/s was about 15 s. In Figure 3, points of flow stress and specimen temperature for a given dwell time are assigned a unique color (see the legend in the figure). For the two tests with impact temperatures just below 300 °C and heating times of 3.5 and 7.5 s (one orange point and one red point in Figure 3), the specimens were heated to a steady temperature of about 360 °C and then allowed to cool prior to impact by delaying the striker impact slightly. With a cooling rate of several hundred degrees per second, the striker delay to achieve the final temperature was of the order of 0.1 s. Experiments above 360 °C for heating times longer than 1 s were heated continually until impact, while experiments heated for 0.3 s were obtained with no dwell time or cooling by using a ramped current pulse. For the temperature range from 250 °C to 350 °C, Figure 3 shows a significant difference in the flow stress between sub-second heating times (green and blue points) and the longer heating times (red and orange points). This temperature range matches that over which precipitate growth is promoted by heat energy (see ref. [3]). It is compelling to think that the effects observed herein are due to the dependency of the size and distribution of the Mg-Si precipitates on heating time. This hypothesis needs to be verified by direct observation of the precipitates. In any case, for applications such as machining, where temperature easily rises several hundred degrees Celsius over a fraction of a second, the elevated-temperature Kolsky compression testing developed for this study may provide thermal softening parameters over applicable, high heating rates. Thermal softening parameters derived from slow heating Kolsky compression tests may not be valid for such problems. Further work will focus on developing a more extensive data set from dynamic compression tests obtained under a wider range of thermal histories, and on the characterization of precipitate growth as a function of thermal history using advanced metallography techniques.

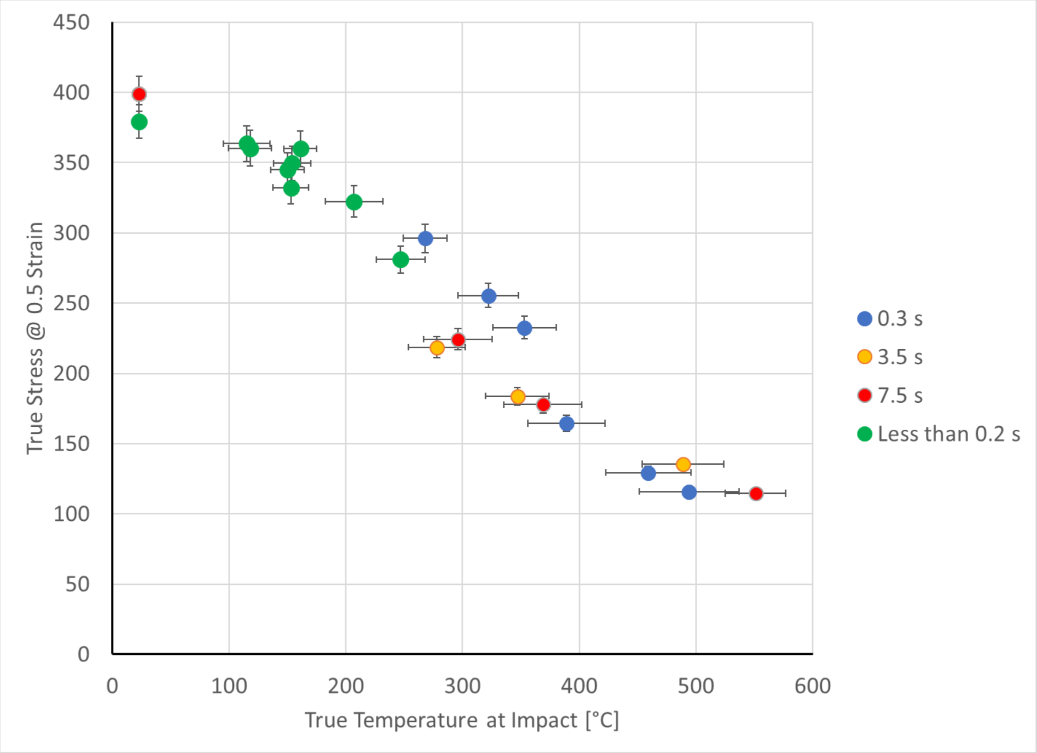


Figure 3. Flow stress of 6061-T6 at several temperatures using the short-pulse-heating method with temperatures determined from the thermal camera. Temperatures are those just prior to impact (within 1.6 ms (=1/614 s)). Error bars on stress are computed from error propagation (95 % confidence interval) and error bars on temperature are computed from thermal gradient measurements obtained during each experiment. The legend shows applied temperature dwell times (heating times).

CONCLUSIONS

At strains of 50 %, strain rates of about 10^3 1/s and near ambient temperature, the flow stress of Al6061-T6 derived from dynamic Kolsky compression tests matches the flow stress derived from machining tests. As the temperature increases, the flow stress decreases. However, differences in thermal softening rates are observed depending on the heating time, owing perhaps to the time-dependent growth of Mg-Si precipitates. The high heating rate, elevated-temperature Kolsky compression test developed for this study offers a unique platform to investigate the effects of time-sensitive material transformations on mechanical response. Future work will expand exploration of time-sensitive material behavior in this model alloy. It will also provide direct observations of the effect of precipitate growth over very short times (of the order of seconds or less) on flow stress, which can be used to develop material models that are better-suited to model machining processes.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Department of Energy, National Nuclear Security Administration under Award Number DE-NA0003222 and the National Institute of Standards and Technology, Measurement, Science & Engineering under Award Number 70NANB19H156. The authors wish to extend special thanks to Pavan Bhavsar and Changlong Cui for providing the data regarding the machining experiments of Al6061-T6, and to Dr. Alkan Donmez of the National Institute of Standards and Technology for facilitating access to EDM.

DISCLAIMERS

(1) This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

(2) This document is an official contribution of the National Institute of Standards and Technology and is not subject to copyright in the USA. The National Institute of Standards and Technology does not recommend or endorse commercial equipment or materials, nor does it imply that the equipment or materials shown are necessarily the best available for the purpose.

REFERENCES

[1] S. Kalpakjian and S. R. Schmid, Manufacturing Processes for Engineering Materials, Upper Saddle River, New Jersey: Pearson, 2007.

[2] J. S. Strenkowski and J. T. Carroll, "A Finite Element Model of Orthogonal Metal Cuting," *Journal of Engineering for Industry,* vol. 107, no. 4, pp. 349-354, 1985.

[3] D. Maisonnette, M. Suery, D. Nelias, P. Chaudet and T. Epicier, "Effects of Heat Treatments on the Microstructure and Mechanical Properties of a 6061 Aluminium Alloy," *Materials Science and Engineering A,* vol. 528, pp. 2718-2724, 2011.

[4] C. Cui, P. Bhavsar, H. Lopez-Hawa, V. Madhavan and W. Moscoso-Kingsley, "Comparison of Flow Stress of Aluminum Alloy 6061-T6 Obtained From Chip-Pulling Orthogonal Cutting and Kolsky Bar Testing," in *SME North American Manufacturing Conference*, Ohio, 2020 (submitted for review).

[5] S. P. Mates, R. Rhorer, B. Timothy and D. Basak, "A Pulse-Heated Kolsky Bar Technique for Measuring the Flow Stress of Metals at High Loading and Heating Rates," *Experimental Mechanics,* vol. 48, pp. 799-807, 2008.

[6] K. Safa and G. Gary, "Displacement Correction for Punching at A Dynamically Loaded Bar End," *International Journal of Impact Engineering,* vol. 37, pp. 371-384, 2010.

[7] A. Manes, L. Peroni, M. Scapin and M. Giglio, "Analysis of strain rate behavior of an Al 6061 T6 alloy," *Procedia Engineering,* vol. 10, pp. 3477-3482, 2011.

A Novel Machine Learning Approach to Estimating KPI and PoC for LTE-LAA-based Spectrum Sharing

Susanna Mosleh[†] [§], Yao Ma[‡], Jacob D. Rezac[‡], and Jason B. Coder[‡]

[†]Associate, Communications Technology Laboratory, National Institute of Standards and Technology, USA
[§]Department of Physics, University of Colorado, Boulder, Colorado, USA
[‡]Communications Technology Laboratory, National Institute of Standards and Technology, USA

***Abstract*—Machine learning (ML) approaches have been extensively exploited to model and to improve wireless communication networks in the past few years. Nonetheless, the estimation of key performance indicators (KPIs) and their uncertainties in Long Term Evolution License Assisted Access (LTE-LAA) based coexistence systems is not adequately addressed. For example, it is not clear if an ML method can accurately predict achievable KPIs (e.g. throughput) and the probability of coexistence (PoC) of LTE-LAA coexistence systems based on partial or no information of MAC and physical layer protocols and parameters. In this paper, we develop a novel ML method by combining a neural network with a logistic regression algorithm to track and estimate KPIs and PoC of coexisting LTE-LAA and wireless local area network (WLAN) links. This ML method can be applied when KPI samples at the base stations (BSs) and access points (APs) are available, without using knowledge of MAC and physical layer parameters. Comparison between the ML and simulation results indicate that the proposed ML method can track the system KPIs and predict the system PoC with good accuracy.**

***Index Terms*—Artificial neural network, LTE-LAA, logistic regression, MAC layer, machine learning, PHY layer, wireless coexistence, WLAN.**

I. Introduction

Wireless communications are tightly integrated in our daily lives. Laptops, tablets, smartphones, and online social networking applications make a level of connectivity to the world available that we have never experienced in the past. This trend continues to dramatically increase wireless network dimensions in terms of subscribers and data throughput [1], especially in the realm of the Internet of Things (IoT). As a consequence, wireless device protocols are beginning to transition from an exclusively-licensed spectrum environment to a shared one, and utilizing the unlicensed spectrum bands seems to be inescapable.

Long Term Evolution (LTE) operating in unlicensed bands, such as license assisted access (LAA), was introduced to improve the spectral efficiency and to help the cellular industry deal with the shortage of the spectrum [2]. However, there are many challenges to overcome in order for multiple networks to constructively share a spectrum. Hence there is a need to accurately evaluate spectrum sharing performance among operators; ensuring the effectiveness of coexistence requires careful consideration.

Recently, artificial intelligence (AI) and machine learning (ML) are having a transformative effect in almost every industry. ML is an important tool for the support of next-generation of wireless device networks. Future 5G and beyond mobile terminals are expected to access the spectral bands using highly developed spectrum learning and inference. However, owing to the involved network topologies, coordination schemes, and the various end-user applications, future networks will be immensely more intricate. Hence, obtaining many optimal key performance indicators (KPIs) might be computationally infeasible or undesirable. Moreover, due to nonlinearity in underlying wireless channels, modeling the end-to-end system's behavior analytically is not easily achieved. It gets even more difficult when it comes to managing networks efficiently in a coexistence scenario where different types of networks need to share a given section of spectrum. ML algorithms can mitigate the underlying unknown non-linearities and can reduce the network complexity so as to be tractable and useful while keeping up ambitious performance goals.

In line with standardization efforts on the evaluation of wireless coexistence [3], [4], in this paper, we develop an ML method to track and estimate the probability of coexistence (PoC) of WLAN and LTE systems in an intelligent and adaptive way. Accurately quantifying the PoC in a given shared spectrum is principal to the evaluation of wireless coexistence, as discussed in the ANSI C63.27 standard [3]. Our proposed method builds up an effective sharing of spectrum, provides an accurate coexistence performance evaluation, and helps to design future radio technologies (e.g., 5G new radio in unlicensed spectrum (5G NR-U) [5]).

The main contribution of this study is to check whether or not ML algorithms can track and provide reliable estimates of KPIs and PoC of coexisting networks. We aim to develop an ML model which provides reliable PoC estimates of various MAC and physical layer parameters, and apply this model to coexistence systems where analytical KPI formulas are not available. These results can support future versions of the ANSI C63.27 standard, and provide insight on developing new ML methods to support KPI uncertainty evaluation in 5G coexistence systems.

We propose a novel PoC estimator to provide an improved assessment of concurrent operation of WLAN and LTE networks in the unlicensed band. Specifically, we take the

Mosleh, Susanna; Ma, Yao; Rezac, Jake; Coder, Jason. "A Novel Machine Learning Approach to Estimating KPI and PoC for LTE-LAA-based Spectrum Sharing." Paper presented at IEEE International Conference on Communications (ICC) Workshop 2020, Dublin, IE. June 07, 2020 - June 11, 2020.

operations of both networks in the MAC and physical layers into account, and employ a novel machine learning algorithm by leveraging neural network with a logistic regression method that utilizes all MAC and physical layers' parameters as inputs (such as contention window size, maximum back-off stage, slot durations, and link signal-to-noise ratio (SNR)) and generates PoC as an output. To do this, we use a neural network to estimate KPIs from input MAC and physical layer parameters. A logistic regression model is then used to estimate PoC from estimated KPIs. The proposed algorithm enables both networks to evaluate wireless coexistence and guarantees a constructive coexistence among operators in an intelligent and a well-planned course of action. It is worth noting that the proposed technique we develop here could be incorporated into many other spectrum sharing systems and we use LTE simply as an example.

The remainder of this paper is organized as follows: Section II describes the system model and assumptions required for our analysis. Section III presents the problem formulation and introduces our proposed intelligent PoC estimator. The impacts of the MAC and physical layer parameters in evaluating a coexistence scenario is also explained in Section III. Simulation results are shown and discussed in Section IV. Finally, in Section V, an overview of the results and some concluding remarks are presented.

II. System Model

Consider a downlink coexistence scenario where two mobile network operators (MNOs) share the same unlicensed bands for operation in an industrial, scientific, and medical (ISM) radio band. Note that we are primarily interested in the operation of cellular base stations in an unlicensed band. However, the LTE base stations may have permission to utilize a licensed band as well. We assume each unlicensed band can be shared between the MNOs in a time sharing fashion. The LAA network consists of L eNodeBs indexed by the set $\mathcal{L} \triangleq \{n_\ell | \ell = 1, 2, \dots, L\}$ while the Wi-Fi network is composed of W APs indexed by the set $\mathcal{W} \triangleq \{n_w | w = 1, 2, \dots, W\}$ APs. Each transmission node $i \in \{\mathcal{L}, \mathcal{W}\}$ serves one single antenna user. The eNodeBs and APs are randomly distributed over a particular area while LAA user equipment (UEs) and Wi-Fi clients are respectively distributed around each eNodeB and AP independently and uniformly. We assume the transmission node i transmits with power p_i and the user association is based on the received power. We also assume (*i*) both Wi-Fi and LAA are in the saturated traffic condition, (*ii*) there is no hidden node problem in the network, *i.e.*, every transmission node i is able to hear one another, (*iii*) a successful transmission happens if only one link transmits at a time, *i.e.*, exclusive channel access (ECA) model is considered, (*iv*) the channel knowledge is ideal (perfect channel sensing among the links), *i.e.*, the only source of packet failure (unsuccessful transmission) is collision, and (*v*) each link is subject to Rayleigh fading and Log-normal shadowing.

Wi-Fi networks utilize a contention-based medium access with a random back-off process, also known as carrier sense multiple access with collision avoidance (CSMA/CA) [6]. In order to discover whether the channel is idle or busy, the station that accesses the medium should sense the channel by performing a clear channel assessment (CCA). The distributed coordination function (DCF) operation progresses if the channel is found out to be idle. Otherwise, the transmitting station refrains from transmitting data until it senses the channel is available. Similarly, LTE uses a listen before talk (LBT) channel access mechanism to maintain fair coexistence with Wi-Fi. Among different LAA-LBT schemes, Category 4 LBT is based on the same Wi-Fi CSMA/CA scheme and is well-suited in a coexistence scenario [2]. Although LTE and Wi-Fi technologies follow the same channel access procedure, they select different carrier sense mechanisms and threshold levels for sensing a channel, and they use different channel contention parameters leading to different unlicensed channel access probabilities which can result in different throughputs.

Our aim is to investigate the feasibility of machine learning algorithms to learn and track system measurement equations (when there is only partial knowledge of MAC/PHY parameters available), or even a system model for which the system equations are not available. To be specific, our goal is to develop a method to map PHY and MAC layers' parameters to PoCs, as depicted in Fig. 1. The PoC here is assessed in terms of the normalized network throughput. Here, we will briefly derive the normalized network throughput of both systems, a quantity that will be used in calculating the PoC later. Conforming with the analytical model in [7], the probability of transmitting a packet by a transmitting node i in a randomly-chosen time slot on an unlicensed channel can be written as

$$p_{\mathrm{tr},i}^{(k)} = \frac{2(1-2p_{c,i}^{(k)})}{(1-2p_{c,i}^{(k)})(1+\mathrm{CW}_{\mathrm{min},i}^{(k)})+p_{c,i}^{(k)}\mathrm{CW}_{\mathrm{min},i}^{(k)}(1-(2p_{c,i}^{(k)})^{m_i^{(k)}})}, \quad (1)$$

where $\mathrm{CW}_{\mathrm{min},i}$ and m_i are the minimum contention window size and the maximum back-off stage of the transmitting node i, respectively, and $p_{c,i}$ is the probability of collision experienced by the i-th transmitting node. The probability of collision experienced by the n_w AP and the n_ℓ eNodeB can be expressed as

$$\begin{aligned} p_{c,n_w} &= 1 - \Big(\prod\nolimits_{\acute{w}\neq w}(1 - p_{\mathrm{tr},n_{\acute{w}}})\Big)\prod\nolimits_{\ell}(1 - p_{\mathrm{tr},n_\ell}), \\ p_{c,n_\ell} &= 1 - \Big(\prod\nolimits_{\acute{\ell}\neq \ell}(1 - p_{\mathrm{tr},n_{\acute{\ell}}})\Big)\prod\nolimits_{w}(1 - p_{\mathrm{tr},n_w}), \end{aligned} \quad (2)$$

respectively, where $\acute{w} = 1, \dots, W$ and $\acute{\ell} = 1, \dots, L$ [8]–[10]. The probability of collision can be split into three parts: the

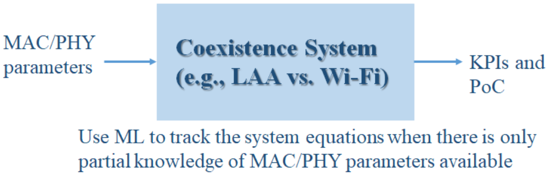

MAC/PHY parameters
Coexistence System (e.g., LAA vs. Wi-Fi)
KPIs and PoC
Use ML to track the system equations when there is only partial knowledge of MAC/PHY parameters available

Fig. 1. System Model

probability of collision due to the collision among the Wi-Fi transmissions, among the LAA transmissions, and between the Wi-Fi and the LAA transmissions, respectively given by

$$p_{c,\mathcal{W}} = (1 - p_{tr,\mathcal{L}})\big[p_{tr,\mathcal{W}} - \sum\nolimits_{w} p_{\text{tr},n_w} \prod\nolimits_{\acute{w}\neq w}(1 - p_{\text{tr},n_{\acute{w}}})\big],$$
$$p_{c,\mathcal{L}} = (1 - p_{tr,\mathcal{W}})\big[p_{tr,\mathcal{L}} - \sum\nolimits_{\ell} p_{\text{tr},n_\ell} \prod\nolimits_{\acute{\ell}\neq \ell}(1 - p_{\text{tr},n_{\acute{\ell}}})\big],$$

and $p_{c,\mathcal{W},\mathcal{L}} = p_{tr,\mathcal{L}} \cdot p_{tr,\mathcal{W}}$. Here $p_{tr,\mathcal{L}} = 1-\prod_{\ell}(1-p_{\text{tr},n_\ell})$ and $p_{tr,\mathcal{W}} = 1 - \prod_{w=1}^{W}(1 - p_{\text{tr},n_w})$ denote the LAA's and Wi-Fi's probability of transmission [8], [9], respectively. Moreover, the probability of a successful transmission by the n_w AP and the n_ℓ eNodeB can be respectively written as [8]–[10]

$$p_{s,n_w} = p_{\text{tr},n_w}\big(\prod\nolimits_{\acute{w}\neq w}(1 - p_{\text{tr},n_{\acute{w}}})\big)\prod\nolimits_{\ell}(1 - p_{\text{tr},n_\ell}),$$
$$p_{s,n_\ell} = p_{\text{tr},n_\ell}\big(\prod\nolimits_{\acute{\ell}\neq \ell}(1 - p_{\text{tr},n_{\acute{\ell}}})\big)\prod\nolimits_{w}(1 - p_{\text{tr},n_w}). \quad (3)$$

Hence, the average length of a time slot can be calculated as

$$T_{\text{avg}} = (1 - p_{tr})\mathbb{E}\{T_{\text{idle}}\} + p_{s,\mathcal{W}}\mathbb{E}\{T_{s,\mathcal{W}}\} + p_{s,\mathcal{L}}\mathbb{E}\{T_{s,\mathcal{L}}\}$$
$$+ p_{c,\mathcal{W}}\mathbb{E}\{T_{c,\mathcal{W}}\} + p_{c,\mathcal{L}}\mathbb{E}\{T_{c,\mathcal{L}}\} + p_{c,\mathcal{W},\mathcal{L}}\mathbb{E}\{T_{c,\mathcal{W},\mathcal{L}}\}$$

where p_{tr} is the probability of occupation of the unlicensed channel, and $p_{s,\mathcal{W}}$ and $p_{s,\mathcal{L}}$ denote the successful transmission probability of the entire Wi-Fi and LAA network, respectively. Moreover, $T_{s,\mathcal{L}}$, $T_{s,\mathcal{W}}$, $T_{c,\mathcal{W}}$, $T_{c,\mathcal{L}}$, and $T_{c,\mathcal{W},\mathcal{L}}$ indicate the time that the channel is being occupied by an LAA successful transmission, a Wi-Fi successful transmission, a collision among the Wi-Fi transmissions, a collision among the LAA transmissions, and a collision between the Wi-Fi and the LAA transmissions, respectively [10]. The network throughput of LAA and Wi-Fi systems as a function of both MAC and physical layers' parameters can be expressed as

$$S_{\mathcal{L}} = p_{s,\mathcal{L}} T_{P,\mathcal{L}} R_{\mathcal{L}} / T_{\text{avg},\mathcal{L}},$$
$$S_{\mathcal{W}} = p_{s,\mathcal{W}} T_{P,\mathcal{W}} R_{\mathcal{W}} / T_{\text{avg},\mathcal{W}}, \quad (4)$$

where $T_{\text{avg},\mathcal{W}}$ ($T_{\text{avg},\mathcal{L}}$) denotes the average time duration to assist a successful transmission in the Wi-Fi (LAA) network, $T_{P,\mathcal{W}}$ ($T_{P,\mathcal{L}}$) indicates the Wi-Fi (LAA) payload duration, and $R_{\mathcal{W}}$ ($R_{\mathcal{L}}$) refers to the Wi-Fi's (LAA's) physical data rate.

Here, we modify the PoC metric described in [11] and define coexistence in terms of the ability to maintain throughput above a certain threshold. Based on the throughput of both LAA and WLAN systems, the PoC metrics that quantify the coexistence performance of these two systems can be calculated as follows

$$\text{PoC}(\eta_{\text{LAA}}, \eta_{\text{WiFi}}) = \mathbb{P}(\text{S}_{\mathcal{L}} > \eta_{\text{LAA}}, \text{S}_{\mathcal{W}} > \eta_{\text{WiFi}}) \quad (5)$$

where equation (5) shows how the joint throughput of Wi-Fi and LAA networks can be mapped to the PoC.

III. Predicting PoC using ML approaches

As mentioned above, an evaluation of the wireless coexistence performance can be determined by the probability of coexistence metric, defined in [11]. This metric will identify the ability of both wireless networks to successfully perform their desired functionality in a given shared spectrum band.

As we discussed earlier, if the physical and MAC layers' parameters are appropriately selected, the throughput of both systems, given by (4), increases, leading to a higher PoC. To be specific, if transmitter i selects the unlicensed spectrum when it is not utilized by the other transmission nodes, then the probability of successful transmission by the i-th node increases, according to (3), leading to a higher throughput. Moreover, if each unlicensed band is selected such that interference is avoided (or at least minimized) among the transmission nodes then there will be a higher SNR, leading to a higher physical data rate (*i.e.,* $R_{\mathcal{L}}$ and $R_{\mathcal{W}}$) and thus higher throughput.

In order to estimate the PoC of Wi-Fi and LAA networks in the unlicensed band, we now develop a machine learning algorithm by leveraging a neural network with a logistic regression method, as depicted in Fig. 2. The proposed model aims to track and estimate KPIs of coexisting LTE-LAA and WLAN links, and evaluate the PoC of these networks. To be specific, the neural network maps the physical and MAC layers input parameters to the KPIs, such as throughput, and the low-cost training logistic regression algorithm conducts a regression analysis to map KPIs to PoC.

A neural network can be thought of as a tracking system used to predict a quantity. It approximates a mapping function from input variables to output variables. In this context, we will use it to track (approximate) the mapping function, i.e., Eq. (4), and determine the LAA and Wi-Fi KPIs based on MAC layer parameters (e.g., contention window size, maximum back-off stages, slot duration) and physical layer characteristics (such as link SNR, link distances, fading parameters). We aim to develop a model which provides reliable approximation of this mapping function of various MAC and physical layer parameters, so we can apply this model to more complex coexistence systems where analytical KPI formulas are not available. The neural network consists of processing nodes organized into three layers, input layer, hidden layer(s), and output layer. These nodes are densely interconnected. In the model that we consider in this paper, known as feed-forward, data moves through the layers in one direction. Each node in a hidden/output layer is connected to several nodes and receives data from them. Moreover, each node in a hidden layer is connected to several nodes and sends data to them. Each connection between nodes is assigned a number named a coefficient or weight. At the training phase, training data is fed to the input layer, it passes through the hidden layer(s), and arrives at the output layer. The data get multiplied by the weights, added together, and go through the activation function of each node. During the training phase, the weights are adjusted until a predefined goal is achieved, such as the mean squared error (MSE) of the training data falling below a pre-set threshold. After this goal is achieved, the trained neural network is applied to test (unseen) data.

Logistic regression can be thought of as a classification algorithm used to assign observations to a discrete set of classes. In this context, we will use it to determine the probability that both LAA and WiFi KPIs are above a user-provided threshold, based on the outputs of the neural network. Logistic regression, when given a set of features, attempts to estimate the probability of success for some function of those features. We use it to estimate the probability of successful coexistence given LAA and Wi-Fi KPIs. In order to solve this prediction problem, we use the gradient descent optimization technique. Let us assume $\mathbf{x}_k$ is the n-dimensional feature vector (the above-mentioned physical and MAC layers' parameters) and y_k is the outcome (PoC of LAA and WLAN networks) of a given test-run $k \in \{1, \ldots, m\}$, where m denotes the total number of observations in the dataset. The feature matrix $\mathbf{X}$ and the outcome vector $\mathbf{y}$ can be written as

$$\mathbf{X}_{n\times m} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{pmatrix},$$

and $\mathbf{y}_{1\times m} = \begin{pmatrix} y_1 & y_2 & \cdots & y_m \end{pmatrix}$. The primary assumption leading to logistic regression is that outcome y_k is Bernoulli-distributed with the success probability π_k. Logistic regression proceeds by estimating the probability of a constructive/destructive coexistence (indicated by the variable $\gamma = 1$ or $\gamma = 0$, respectively) given the training set $\mathbf{x}_k$, *i.e.,* $\hat{y}_k = \mathbb{P}(y_k = \gamma|\mathbf{x}_k) = \pi_k^{\gamma}(1-\pi_k)^{1-\gamma}$.

In order to estimate this probability using the logistic regression method, we need to find a hypothesis $h_{\boldsymbol{\theta}_k}(\mathbf{x}_k) = \hat{y}_k$. The goal is to learn the optimum value of the regression coefficients $\boldsymbol{\theta}_k$ in the sense that $\hat{y}_k$ is approximately equal to the test target y_k. $\boldsymbol{\theta}_k$ is the set of weights corresponding to n features and the bias. In order to learn these weights, we need to define a cost function. A cost function is an estimator of how well our model predicts the known output. This cost function will be used to train the logistic regression model (prediction function) that could predict the PoC in unlicensed band. The logistic regression model can be given as

$$\hat{\mathbf{y}} = \mathrm{S}(\mathrm{diag}(\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{X})), \tag{6}$$

where $\hat{\mathbf{y}}_{1\times m} \triangleq \begin{pmatrix} \hat{y}_1 & \cdots & \hat{y}_m \end{pmatrix}$, $diag(\mathbf{A})$ returns a vector of the main diagonal elements of $\mathbf{A}$, $\boldsymbol{\Theta}_{n\times m}$ is the weight matrix, subscript T is the transpose operator, and $S(z) = \exp(z)/(1+\exp(z))$ is the so-called sigmoid (logistic) function. The sigmoid function $S(z)$ introduces non-linearity to the model and maps predicted values to probabilities. Then, the hypothesis of logistic regression for the training pair $(\mathbf{x}_k, y_k)$ can be written as

$$\hat{y}_k(\mathbf{x}_k) = h_{\boldsymbol{\theta}_k}(\mathbf{x}_k) = \frac{1}{1+e^{-\boldsymbol{\theta}_k^T\mathbf{x}_k}}. \tag{7}$$

In order to calculate the weight matrix $\boldsymbol{\Theta}$, a cost function is needed for optimization. Cost functions are usually defined as MSE functions. However, it is known that when using this cost function, the optimization problem turns out to be non-convex and has many local minimums [12, Chapter 3]. Hence, in this paper we use a cost function called "cross-entropy", also known as log loss function, for each pair of training samples, *i.e.,* $(\mathbf{x}_k, y_k)$, as follows [13, Chapter 5]

$$\mathcal{L}(y_k, \hat{y}_k) = -y_k \log(h_{\boldsymbol{\theta_k}}(\mathbf{x}_k)) - (1-y_k)\log(1-h_{\boldsymbol{\theta_k}}(\mathbf{x}_k)),$$

which plays the same role as the MSE function, but now the optimization problem becomes convex in $\boldsymbol{\theta}$. This cost function turns the optimization problem into a convex one which is much easier to solve using standard computational techniques. $\mathcal{L}(y_k, \hat{y}_k)$ shows how well the prediction is in a single training example. The cost function of all training samples used in the logistic regression algorithm can be expressed as

$$\begin{aligned}\mathcal{J}(\boldsymbol{\Theta}) &= \frac{1}{m}\sum_{k=1}^{m}\mathcal{L}(y_k, \hat{y}_k) \\ &= \frac{1}{m}(-\mathbf{y}^T \log\hat{\mathbf{y}} - (1-\mathbf{y})^T\log(1-\hat{\mathbf{y}})).\end{aligned} \tag{8}$$

By minimizing this cost function with respect to $\boldsymbol{\theta_k}$, the optimum value of the weight matrix $\boldsymbol{\Theta}$ can be found using following optimization

$$\min_{\boldsymbol{\Theta}} \quad \mathcal{J}(\boldsymbol{\Theta}). \tag{9}$$

In order to minimize the cost function $\mathcal{J}(\boldsymbol{\Theta})$ we apply the gradient descent method, which is the most popular approach to iteratively minimize the cost function. The update equation for the k-th observation in the data set can be written as

$$\begin{aligned}\boldsymbol{\theta}_k &= \boldsymbol{\theta}_k - \alpha\nabla_{\boldsymbol{\theta}_k}\mathcal{J}(\boldsymbol{\Theta}) \\ &= \boldsymbol{\theta}_k + \alpha(y_k - h_{\boldsymbol{\theta}_k}(\mathbf{x}_k))\mathbf{x}_k,\end{aligned} \tag{10}$$

where $\nabla\mathcal{J}$ denotes the gradient of the function $\mathcal{J}$, and $0 \leq \alpha \leq 1$ is the step size, also known as the learning rate, and determines how fast the learning happens. As is typically the case in learning algorithms, selecting α requires some care. A small value of α results in a long learning process (which could be detrimental in practice), while a large value of α could cause bouncing around the optimum point.

After calculating the optimal weight matrix $\boldsymbol{\Theta}$, the new label of an unseen sample can be estimated by using (7). To map the estimated label to a discrete class (constructive/destructive coexistence), the predefined threshold value η_{PoC} is selected above which we will classify values as constructive coexistence (high PoC) and below which we classify values as destructive coexistence (low PoC).

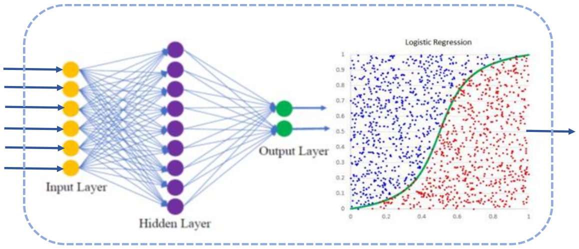


Fig. 2. Proposed Model

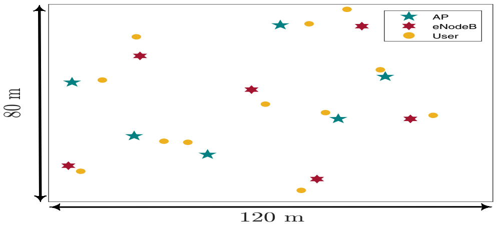


Fig. 3. Simulation Layout

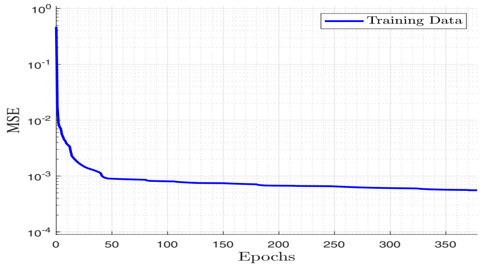


Fig. 4. Mean squared error on the training data in the NN

TABLE I
MAC Layer Parameters

Parameter	value
LAA's packet payload duration	2 ms
Wi-Fi's packet payload duration	1 ms
MAC header	272 bits
PHY header	128 bits
ACK	112 bits + PHY header
SIFS	16 μs
DIFS	34 μs
Idle slot time	9 μs
Wi-Fi contention window size	16
LAA contention window size	16
Wi-Fi maximum backoff stage	6
LAA maximum backoff stage	3

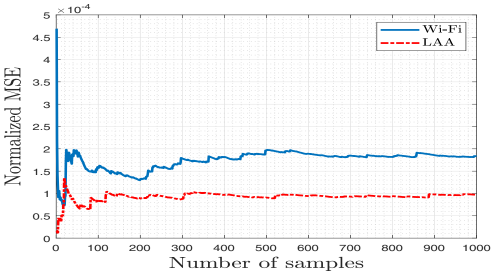


Fig. 5. Normalized mean squared error on the test data in the NN

IV. Simulation Results and Discussions

We evaluate the performance of the proposed algorithm in a coexistence scenario. We simulate a scenario in which 6 eNodeBs compete for an unlicensed channel with 6 APs. All transmitters are randomly distributed over an area of size 120×80 m^2 with a minimum distance of 20 meters, as shown in Fig. 3. All UEs and Wi-Fi clients are independently and uniformly distributed around each eNodeB and AP, respectively. We consider one UE (Wi-Fi client) per eNodeB (AP). Each UE (Wi-Fi client) is assigned to the eNodeB (AP) that provides it with the highest received power. The antenna height of the transmission nodes and users are 6 meters and 1.5 meters, respectively. The carrier frequency is 5 GHz and the bandwidth of each channel is 20 MHz. The path-loss and shadowing between transmission nodes and users are generated following [14] for the indoor scenario. The transmit power at each transmission node is fixed to 23 dBm while the noise figure and the thermal noise level at each user is set to 9 dB and -174 dBm/Hz, respectively [14]. Moreover, we assume the omni-directional antenna pattern with a 0 dBi antenna gain.

According to this geometry and propagation model, we compute the SNR of each link. Given the MAC layer parameters in Table I, we select a sample set of input parameters, and train the neural network to generate output KPIs of both LAA and WLAN networks. The data used for this simulation consists of 1000 feature vectors. Only 30% of the data are used for training and the rest are considered for test. Here, we consider a feedforward neural network consisting of one input layer with 16 nodes (Wi-Fi and LAA contention window sizes, Wi-Fi and LAA backoff stages, and 12 link SNRs), one hidden layer with 16 neurons, and one output layer with two nodes (LAA's and Wi-Fi's throughput). The network is trained and converges quickly, as shown in Fig. 4. After training the network, we applied the trained neural network on unseen (test) data. In order to evaluate the trained network, we calculated the MSE as an average of the squared error $(\mathbf{y}-\hat{\mathbf{y}})^2$, where $\mathbf{y}$ is the KPI values calculated using equation (4) and $\hat{\mathbf{y}}$ is the output of the trained neural network. Comparing the neural network's outputs with the analytical results, the MSE of the test data is equal to 0.0043. The small value of MSE on test data indicates that the neural network tracks the mapping function (system equations) well. We also plot the normalized MSE of both Wi-Fi and LAA KPIs in Fig. 5.

Having the LAA and WLAN throughput pairs as the outputs of neural network, now we map the KPI pair to PoC using the trained logistic regression model and then compare the results with the theoretical one found by Eq. (5). In order to train the logistic regression model, given the predefined thresholds η_{LAA} and η_{WiFi}, we first plot the probability of satisfactory quality of service (QoS) of LAA and Wi-Fi networks, *i.e.*, $\mathbb{P}(S_{\mathcal{L}} > \eta_{\text{LAA}})$ and $\mathbb{P}(S_{\mathcal{W}} > \eta_{\text{WiFi}})$, in Fig. 6 and Fig. 7, respectively. It is observed that the analytical and neural network results follow the same trend. As expected, by increasing the thresholds, the probability of satisfactory QoS of each system decreases. Moreover, the probability of coexistence versus the Wi-Fi's and LAA's throughput threshold is plotted in Fig. 8. The goal is to train the logistic regression model to accurately estimate PoC from estimated throughput and enable both networks to evaluate the wireless coexistence. We pass the test data to the trained logistic regression network.

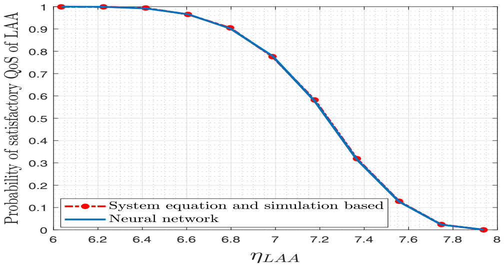


Fig. 6. Probability of satisfactory QoS of LAA versus different thresholds

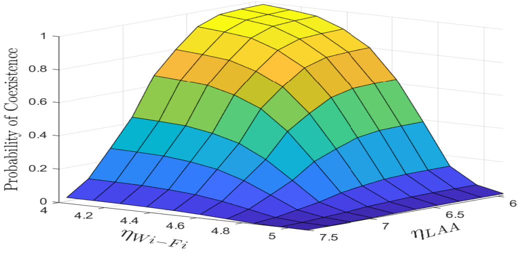


Fig. 8. PoC versus different Wi-Fi's and LAA's throughput threshold

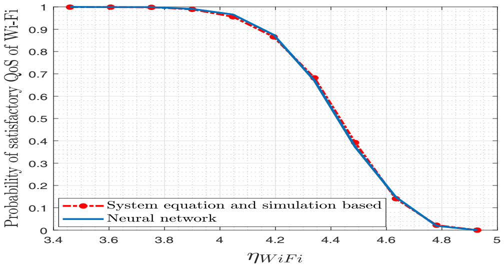


Fig. 7. Probability of satisfactory QoS of Wi-Fi versus different thresholds

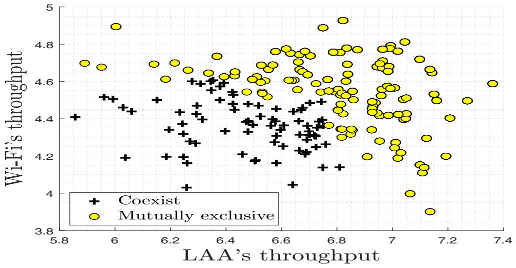


Fig. 9. Labeling Wi-Fi's and LAA's throughput pair using the trained LR

The logistic regression network labeled the unseen data as they can/cannot coexist with each other. Fig. 9 shows that by knowing the throughput of two networks we are able to decide whether or not these two networks can coexist with each other. Moreover, we compare the PoC calculated by equation (5) with the PoC output of the trained logistic regression network and compute the MSE. The MSE on the unseen test data is 0.0658. Furthermore, the logistic regression accuracy, which is defined as percentage of correct predictions, on the training and test data is calculated and given as 84.466% and 84.046%, respectively. In future work we develop further enhanced classification schemes to improve the accuracy.

V. Conclusion

In this paper, we have proposed a machine learning method to accurately track and estimate coexistence performance and its uncertainty between LTE-LAA and WLAN networks in unlicensed bands. The proposed method can work without knowledge of MAC and physical layer protocols and parameters of the system (except the KPI samples). We have also developed a system equation and simulation-based scheme to train and validate the performance of our method. Comparison of the proposed ML method and simulation results has demonstrated that our method can achieve a perfect KPI (i.e., throughput) tracking and a good PoC estimation performance. The proposed method can be extended to the cases where analytical results are not available. In future work, we will develop further enhanced ML methods for KPI and an uncertainty estimation, and generalize our proposed approach to more challenging coexistence scenarios.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 20172022 White Paper," Cisco Systems, Febuary 2019.

[2] "Study on licensed-assisted access to unlicensed spectrum," 3GPP TR. 36.889 v13.0.0., Jun. 2015.

[3] "American National Standard for Evaluation of Wireless Coexistence," *in ANSI C63.27-2017*, pp. 1-77, May 2017.

[4] "IEEE Standard Definitions and Concepts for Dynamic Spectrum Access: Terminology Relating to Emerging Wireless Networks, System Functionality, and Spectrum Management," *in IEEE Std 1900.1-2008*, pp. 1-62, Oct. 2008.

[5] S. Verma and S. Adhikari, "3GPP RAN1 status on LAA and NR Unlicensed," *IEEE 802.11-18/0542r0*, Mar. 2018.

[6] E. Perahia and R. Stacey, "Next Generation Wireless LANs: 802.11n and 802.11ac," Cambridge, U.K.: Cambridge Univ. Press, June 2013.

[7] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," IEEE J. Sel. Areas Commun., vol. 18, no. 3, pp. 535-547, March 2000.

[8] S. Mosleh, Y. Ma, J. B. Coder, E. Perrins, and L. Liu, "Enhancing LAA co-existence using MIMO under imperfect sensing", *IEEE GLOBECOM Wkshps*, pp. 1 - 6, Waikoloa, HI, Dec. 2019.

[9] S. Mosleh, Y. Ma, J. D. Rezac, J. B. Coder, "Dynamic spectrum access with reinforcement learning for unlicensed access in 5G and beyond", accepted for publication in the *IEEE VTC Spring*, 2020.

[10] Y. Ma and D. G. Kuester, "MAC-layer coexistence analysis of LTE and WLAN systems via listen-before-talk," *IEEE CCNC*, pp. 534 - 541, Jan. 2017.

[11] Y. Ma, W. Young, E. Anderson, and J. Coder, "Probability of coexistence of LTE-LAA and WLAN systems based on delay constraints", *IEEE ICCCN*, pp. 1 - 9, Aug. 2018.

[12] M. Nielsen, "Neural Networks and Deep Learning," Determination Press, 2015.

[13] R. Klimberg, and B. D. McCullough, "Fundamentals of Predictive Analytics with JMP," SAS Institute, April 2012.

[14] "3GPP TSG RAN; Study on NR-based access to Unlicensed Spectrum; (Release 16)," 3GPP TR 38.889 V16.0.0, December 2018.

Contributions to Acoustic Loss in Langasite, Langatate, and Catangasite Resonators at High Temperatures

Ward L. Johnson
Applied Chemicals and Materials Division, National Institute of Standards and Technology, 325 Broadway St., MS 647, Boulder, CO 80305, USA
wjohnson@boulder.nist.gov

Summary:

Piezoelectric crystals in the langasite family can serve as a basis for resonant acoustic sensors that operate at temperatures exceeding the range of conventional piezoelectric resonators, but their performance is limited by acoustic loss at elevated temperatures. This paper presents an overview of what is currently known and not known about physical contributions to temperature-dependent acoustic loss in langasite, langatate, and catangasite [1].

Keywords: acoustic loss, attenuation, high temperature, langasite, langatate, catangasite, piezoelectric resonators

Background

Applications of resonant piezoelectric sensors have traditionally been limited to temperatures below several hundred degrees Celsius, because phase transitions and/or material degradation occur at higher temperatures in common commercially available piezoelectric materials [2]. However, over the past four decades, substantial research has focused on growing, characterizing, and optimizing innovative piezoelectric crystals that can be used in resonant sensors at temperatures exceeding 1000 K [3]. These piezoelectrics include crystals with the structure of langasite ($La_3Ga_5SiO_{14}$, LGS).

A critical factor in the performance of resonators is the quality factor Q and corresponding acoustic loss Q^{-1}. This paper is focused on contributions to Q^{-1} in LGS and two piezoelectrics with similar crystal structure, langatate ($La_3Ga_{5.5}Ta_{0.5}O_{14}$, LGT) and catangasite ($Ca_3TaGa_3Si_2O_{14}$, CTGS).

LGS

Figure 1 shows measurements of Q^{-1} acquired at the National Institute of Standards and Technology (NIST, U.S.A.) and Clausthal University of Technology (TUC, Germany) on two Y-cut crystals grown by different manufacturers [2,4]. The combined data in this figure extend over an exceptionally broad temperature range and thereby provide a unique illustration of typical contributions to the temperature-dependent loss in LGS and other crystals in the langasite family. The NIST measurements were performed with noncontacting electrodes in vacuum, and the TUC measurements were performed with Pt surface electrodes in air.

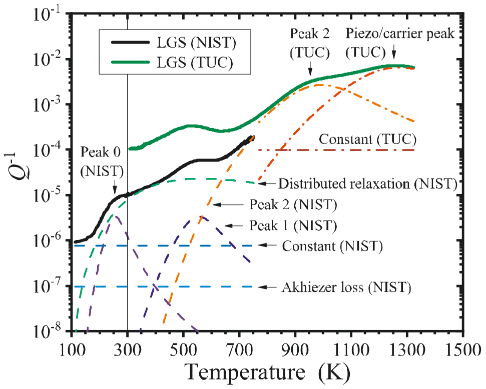


Fig. 1. Acoustic loss Q^{-1} of two LGS crystals measured at NIST and TUC. Resonant frequencies were 6.05 MHz (NIST) and 5.0 MHz (TUC) at ambient temperature. Dashed and dashed-dotted lines are contributions to Q^{-1} of the NIST and TUC crystals, respectively, determined from least-squares fits.

The three peaks below 1100 K in Fig. 1 are consistent with anelastic point-defect relaxations with a characteristic Debye dependence on temperature T and angular frequency ω [5]:

$$Q^{-1} = (\Delta/T)\omega\tau/(1+\omega^2\tau^2) , \qquad (1)$$

where the relaxation strength Δ is proportional to defect concentration. The relaxation time τ has an Arrhenius dependence on T,

$$\tau = \gamma\exp(U/kT) , \qquad (2)$$

where U is an activation energy on the order of an electron volt and k is Boltzman's constant.

Another peak appears in the TUC data near 1260 K, and this is associated with piezoelectric/carrier relaxation, involving the motion of charge carriers in acoustically generated piezoelectric fields [2,4]. In LGS, this peak is expected to have Debye form, but without T in the pre-factor [4]. An additional contribution in all crystals is intrinsic loss from phonon-phonon interactions (Akhiezer loss) [6], which is expected to be proportional to ω and weakly dependent on T above 100 K [4].

The data from NIST in Fig. 1 (112 K to 752 K) were fit to a function that includes three point-defect relaxations, Akhiezer loss approximated as independent of T, a constant background, and a broad relaxation consisting of a continuous set of Debye functions with a log-normal distribution of activation energies. The last term was found necessary to accurately fit the data [4]. Piezoelectric/carrier loss was not included, because initial analysis indicated it to be a minor contribution at all measured temperatures. The results of this fit at 6 MHz, simultaneously performed on data from two additional harmonics, are shown in the figure.

A fit of the TUC data (Fig. 1) accurately matches the data without including a distributed relaxation term. To simplify the figure, contributions to this fit are only plotted above the range of the NIST data and Peak 1 is not included. The piezoelectric peak position is consistent with predictions based on the measured temperature-dependent conductivity and dielectric constant [7]. The constant term in the fit is $1.0\text{x}10^{-4}$, two orders of magnitude greater than the constant in the fit of the NIST data. This difference may be due to greater mechanical contact and/or anelasticity in the Pt electrodes.

LGT and CTGS

Similar point-defect and piezoelectric/carrier contributions have been reported in LGT and CTGS at resonant frequencies in the low megahertz range [4,7,8]. Two point-defect peaks are typically observed between 300 K and 1000 K. Lower conductivity and corresponding τ_c of CTGS, relative to LGS, lead to lower piezoelectric carrier loss over the measured ranges of temperature, since the peak maximum is shifted to higher temperatures [7].

In an LGT crystal measured at NIST, no broad temperature-dependent contribution similar to that observed in LGS (NIST, Fig. 1) was detected [4]. The fact that dislocation density was much lower in this LGT crystal supports the hypothesis that the physical mechanism responsible for this contribution in LGS involves anelastic motion of kinks in dislocations [4].

Suhak *et al*. [7] found evidence for a broadly temperature-dependent background loss in Y-cut CTGS and employed a simpler Arrhenius form to fit this background. They also found the constant term in the fitting function to be much greater for a CTGS crystal with Pt electrodes than for a different CTGS crystal without electrodes.

Conclusions

LGS, LGT, and CTGS display similar features in the temperature dependence of acoustic loss, although the magnitudes and peak positions of loss contributions vary. The general form of the piezoelectric/carrier contribution and its dependence on electrical conductivity is well understood. However, specific defects responsible for point-defect peaks have not been identified. Such identification may be less than straightforward because correlations of peaks with impurity concentrations can be indirect. In particular, defect symmetries and associated relaxations can depend on charge states that change with dopant levels. The physical mechanisms responsible for observed contributions with broad temperature dependence are also not established. The combination of results from multiple crystals suggests that these mechanisms include effects that are both internal and external to the piezoelectric material.

[1] This manuscript is a contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.

[2] Fritze, H., High-temperature piezoelectric crystals and devices, *J. Electroceram.* 26, 122-161 (2011); doi: 10.1007/s10832-011-9639-6

[3] Johnson, W., Acoustic and Electrical Properties of Piezoelectric Materials for High-Temperature Sensing Applications, *Proc. SENSOR 2015*, 384-389 (2015); doi: 10.5162/sensor2015/C3.1

[4] Johnson, W., Kim, S. A., Uda, S., and Rivenbark, C. F., Contributions to anelasticity in langasite and langatate, J. Appl. Phys. 110, 123528 (2011); doi: 10.1063/1.3672443

[5] A. S. Nowick and B. S. Berry, *Anelastic Relaxation in Crystalline Solids* (Academic, NY, 1972).

[6] Akhiezer, A., On the absorption of sound in solids, *J. Phys. USSR 1*, 277-287 (1939).

[7] Suhak et al., Electromechanical properties and charge transport of $Ca_3TaGa_3Si_2O_{14}$, Solid State Ionics 317, 221-228 (2018); doi: 10.1016/j.ssi.2018.01.032

[8] Johnson, W. L., High-Temperature Electroacoustic Characterization of Y-Cut and Singly-Rotated $Ca_3TaGa_3Si_2O_{14}$ Resonators, *IEEE T. Ultrason. Ferr.* 61, 1433-1441 (2014); doi: 10.1109/TUFFC.2014.3052

Ohms Law Low-current Calibration System for Ionization Chambers

Dean G. Jarrett*, Shamith U. Payagala*, Ryan Fitzgerald*, Denis E. Bergeron*, Jeffrey T. Cessna*, Charles J. Waduwarage Perera†*, and Neil M. Zimmerman*

*National Institute of Standards and Technology, 100 Bureau Drive, Stop 8171, Gaithersburg, MD, 20899, USA
dean.jarrett@nist.gov

†Department of Chemistry and Biochemistry, University of Maryland, College Park, MD, 20742, USA

***Abstract* — A system for the calibration of electrometers that measure currents from ionization chambers is described. The calibration system uses a 1 GΩ standard resistor in series with a stable voltage source to generate calibration currents from 1 pA to 20 nA, traceable to quantum voltage and resistance standards through an 8-1/2 digit digital voltmeter and the standard resistor. Expanded uncertainties (k = 2) of 100 × 10^{-6} or better for the calibration of the electrometer is needed for the ionization chamber measurements. The electrometer calibration was verified by current measurements following the decay of the short-lived radionuclide ^{18}F in a re-entrant ionization chamber.**

***Index Terms* — Electrometer, ionization chamber, quantum electrical standards, standard resistor, small current measurement.**

I. Introduction

For decades, radionuclide metrology has relied on re-entrant ionization chambers for comparing and maintaining secondary measurement standards and measuring half-lives in support of nuclear medicine, nuclear power, and other fields [1]-[2]. The output of the ionization chambers is a small current in the range of 10^{-13} A to 10^{-8} A, depending upon the radionuclide and activity. The ionization chamber is sealed vessel containing pressurized gas, typically nitrogen at 1 MPa. When a radionuclide source is placed in the well in the center of the ionization chamber, the gas is ionized by gamma rays from the decaying radionuclide and the freed charge is collected by applying high voltage (typically 500 V) across the chamber. The output current is proportional to the activity of the sample.

Linearity and stability over the 5 decades of currents generated by ionization chambers is a major uncertainty component and has been addressed for many years by using long-lived sealed radioactive sources to generate reference currents as standards to measure other sources. Highly active reference sources, such as ^{226}Ra, have been relied upon in the past. But reduced availability of the reference radioactive sources in the future, the aging of electronics used for the ionization current measurements, and improvements in electrical metrology for low current measurements have made the linking of ionization chamber measurements to the quantum SI through electrical units an attractive alternative [3].

One approach, described here, is to use a calibrated high value standard resistor, a stable voltage source, and an 8-1/2 digit digital voltmeter (DVM) to generate currents which are traceable to quantum resistance and voltage standards. This measurement system provides a cost-effective method to routinely calibrate the ionization chamber electrometer in the laboratory with an expanded uncertainty (k = 2) of 100 × 10^{-6} or better for the 20 pA to 20 nA ranges.

II. Measurement System

The main components of the measurement system are the standard resistor, the 8-1/2 digit DVM, the stable voltage source, and a switch. These instruments, along with a temperature monitor, cables, interfaces, and optical isolators have been packaged in a small portable instrument rack; allowing the system to be easily moved.

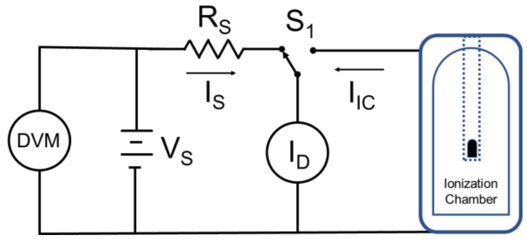


Fig. 1. Circuit diagram for calibration of electrometer used with an ionization chamber. Switch S_1 connects the electrometer I_D (measuring currents I_S and I_{IC}) to the ionization chamber or the voltage source V_S, DVM, and standard resistor R_S.

The currents are generated with a standard resistor and a voltage source in series, where Ohm's Law defines the calibration current. Four ranges were identified (20 pA, 200 pA, 2 nA and 20 nA) with a goal of 100 × 10^{-6} or better expanded relative uncertainty (k = 2). The standard resistor was calibrated with traceability to the quantum Hall resistance (QHR); a DVM, to measure the voltage source, was calibrated with traceability to the Josephson voltage standard (JVS). A 1 GΩ standard resistor was chosen with a voltage source of 0.02 V to 20 V to generate the required currents to calibrate the electrometer at full scale for each range. The relative standard uncertainty for calibration of the 1 GΩ standard resistor was 5 × 10^{-6} and the relative standard uncertainty for calibration of the DVM was 10 × 10^{-6}. Figure 1 shows the ionization chamber and the electrical calibration system.

The standard resistor has been fabricated and calibrated in the NIST resistance laboratory [4]. The resistance element has been

Jarrett, Dean G.; Payagala, Shamith; Fitzgerald, Ryan P.; Bergeron, Denis E.; Cessna, Jeffrey T.; Waduwarage Perera, Charles; Zimmerman, Neil M. "Ohms Law Low-current Calibration System for Ionization Chambers." Paper presented at 2020 Conference on Precision Electromagnetic Measurements(CPEM 2020), Denver, CO, US. August 24, 2020 - August 28, 2020.

heat treated and hermetically sealed for long-term stability and characterized for temperature and voltage dependence. The drift rate has also been determined to be less than 1 x 10^{-5}/year so the calibration interval for the resistor can be greater than one year. For this application, the standard resistor was rack mounted in an insulated enclosure at the bottom of the rack to minimize heating from the other instruments. A thermistor was permanently mounted on the sealed resistor canister (inside the insulated enclosure) so the resistance value can be corrected for temperature.

The voltage burden of the electrometer pushes uncertainty estimates above the target uncertainty at the lower current ranges of 20 pA and 200 pA [5]-[6]. The voltage burden has been found to be stable and has a Gaussian distribution during short-term intervals of 5 minutes with a relative standard deviation of 2.9 x 10^{-4}, indicating that the electrometer voltage burden is stable for a calibrate – measure – calibrate procedure. A switch is used to connect the electrometer to either the ion chamber or the electrical calibration system.

III. Ionization Chamber Measurements

The calibration of two electrometers verified that the instruments were within the manufacturer specification and showed similar results. Figure 2 shows the difference between the measured current and the calculated calibration current across five decades of current from 2 pA to 20 nA for an electrometer used with the ionization chamber. The solid lines show the manufacturer's specification for the electrometer. Data points are the difference between the measured net current and the calibration current based on the standard resistor and the voltage source.

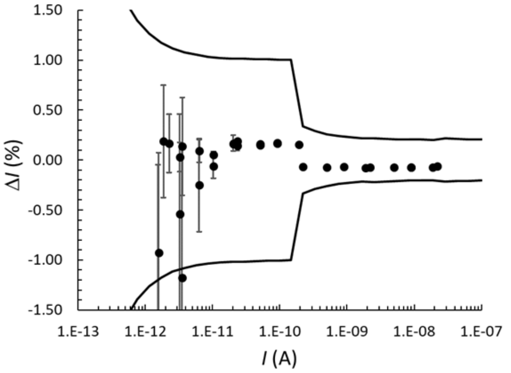


Fig. 2. Electrometer calibration within the manufacturer specification. Error bars are the standard deviation of the repeated measurements.

A test of the calibration system was made using the short-lived radioactive source ^{18}F over four decades of current as shown in Fig. 3. After calibration, the residuals from fits to the exponential decay of the current were reduced both in magnitude and in range-change artifacts.

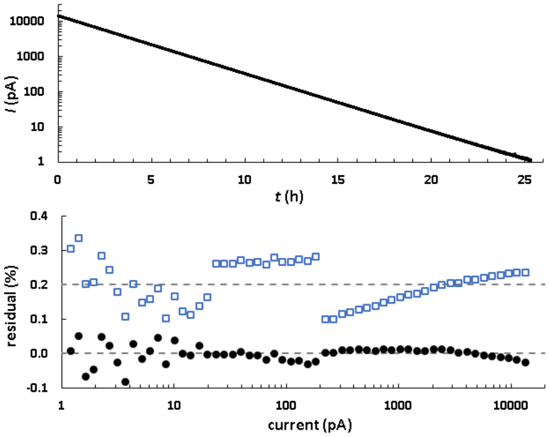


Fig. 3. Upper panel shows decay of ^{18}F from 10 nA to 1 pA. Lower panel shows fit residuals before (□) and after (●) calibration correction. For clarity, the before residual has been offset by + 0.2 %.

IV. Conclusion

A calibration system to provide ionization chamber measurements traceable to electrical units in the quantum SI is described. The calibration currents are sourced from a 1 GΩ standard resistor and a voltage source, with a digital voltmeter measuring the sourced voltage. Calibration of the ionization camber electrometer with relative uncertainty better than 0.1 % over the range 10^{-12} A to 10^{-8} A has been demonstrated, improving the measurement uncertainties and reducing reliance on high activity radioactive sources in the regular calibration workflow. Testing with the short-lived radionuclide ^{18}F was done to verify the calibration over a four-decade range.

References

[1] M. N. Amiot, V. Christie, R. Fitzgerald, F. Juget, C. Michotte, A. Pearce, G. Ratel, B. E. Zimmerman, "Uncertainty evaluation in activity measurements using ionization chambers", *Metrologia,* 52, S108-S122, 2015.

[2] H. Shrader, "Half-life measurements with ionization chambers—A study of systematic effects and results", *Applied Radiation and Isotopes,* 60, 317-323, 2004.

[3] R. Fitzgerald, D. E. Bergeron, S. P. Giblin, D. G. Jarrett, S. M. Judge, C. Michotte, H. Scherer, N. M. Zimmerman, "The Next Generation of Current Measurements for Ionization Chambers", submitted to *Applied Radiation and Isotopes*, 2019.

[4] R. F. Dziuba, D. G. Jarrett, L. L. Scott, A. J. Secula, "Fabrication of High-Value Standard Resistors", *IEEE Trans. on Instrum. and Meas.*, 48, 2, 333-337, 1999.

[5] L. Callegaro, P. P. Capra, V. D'Elia, F. Galliana, "Generation of Reference DC Currents at 1 nA Level with the Capacitance-charging Method", *IEEE Trans. on Instrum. and Meas,* 63, 1779-1782, 2014.

[6] Low level measurements handbook - 7th edition: "Precision DC Current, Voltage, and Resistance Measurements", Keithley Instruments, A Tektronix Company, 2016.

Deterministic Intra-Vehicle Communications: Timing and Synchronization

Hamid Gharavi and Bin Hu
Advanced Network Technologies
National Institute of Standards and Technology
Gaithersburg, USA
Emails: [Gharavi, bhu]@nist.gov

Abstract— **As we power through to the future, in-vehicle communications' reliance on speed is becoming a challenging predicament. This is mainly due to the ever-increasing number of electronic control units (ECUs), which will continue to drain network capacity, hence further increasing bandwidth demand. For a wired network, a tradeoff between bandwidth requirement, reliability, and cost-effectiveness has been our main motivation in developing a high-speed network architecture that is based on the integration of two time-triggered protocols namely; Time Triggered Ethernet (TT-E) and Time Triggered Controller Area Network (TT-CAN). Therefore, as a visible example of an Internet of Vehicles technology, we present a time triggered communication-based network architecture. The new architecture can provide scalable integration of advanced functionalities, while maintaining safety and high reliability. To comply with the bandwidth requirement, we consider high-speed TT-Ethernet as the main bus (i.e., backbone network) where sub-networks can use more cost-effective and lower bandwidth TT-CAN to communicate with other entities in the network via a gateway. The main challenge in the proposed network architecture has been to resolve interoperability between two entirely different time-triggered protocols, especially in terms of timing and synchronization. With that in mind, we explore the main key drivers of the proposed architecture, which are bandwidth, reliability, and timeliness.**

Keywords—Internet of Vehicles (IoV), Time Triggered Ethernet (TT-E), Time Triggered Controller Area Network (TT-CAN), Electronic Control Units (ECUs), Intra Vehicle Communications

I. INTRODUCTION

A rapid growth in electric vehicles and Advanced Driver Assistant Systems (ADAS) is transforming future cars into large mobile data centers. The most important example of such a transformation is the self-driving car technology, which requires a large number of smart devices and sensors that can guarantee much better safety than manual driving. This, together with the increasing scale and scope of data generated from a vehicle with multiple ECUs, represents a tremendous challenge for future intelligent transportation under the auspices of the Internet of Vehicle (IoV).

Currently, vehicle electronics consist of several "sub-systems" or "domains" where each has its own control units, such as mechanical, electrical, or computer controls. The data generated by these sub-systems can vary with respect to bandwidth, reliability, and latency requirements. There are several advanced intra-vehicle communications standards, such as TT-CAN [1] and FlexRay [2, 3]. In particular, TT-CAN, which is an extended version of CAN [4], has been one of the most cost-effective communications deployed in today's vehicles [4]. While TT-CAN can fulfill the short term need for time-critical in-vehicle communications, its limited bandwidth is insufficient to respond to the ever increasing demand for more bandwidth. Such a demand has been further intensified in recent years by the introduction of more bandwidth hungry services, such as infotainment, camera-based Advanced Driver Assistance Systems (ADAS), advanced Laser detection and ranging (LADAR) scanners for crash detection and prevention [5], as well as future advanced 3-dimensional range video [6].

As a long-term solution, Ethernet technology, which is capable of providing a much higher bandwidth, embraces a logical transition towards IoV [7-9]. Currently, there are a few Ethernet-based protocol standards for in-vehicle communications [10]. While these protocols have undergone massive advancements, the reliability of the network to respond to emergency situations in a timely manner has been the main requirement.

Therefore, to keep up with the increasing integration of multiple sub-networks and cope with the staggering amount of video streaming for both safety and entertainment, in-vehicle communication would require not only faster, but highly reliable Ethernet-based networks, such as Time-triggered Ethernet (TT-E) [11]. On the other hand, transitioning from non-Ethernet-based TT-CAN to TT-E is a huge undertaking in terms of cost as most existing in-vehicle components are generally developed for CAN. Therefore, as a comprise between cost and capacity, in this paper we present a new network architecture, which is based on two time-triggered protocols. The main objective of integrating the two protocols is to expand the bandwidth and yet reduce the overall cost as much as possible. The main challenge, however, has been to achieve interoperability between the two time-triggered protocols. The proposed architecture provides a scalable integration of both protocols with advanced functionalities, while ensuring safety and reliability.

After a brief overview of the existing protocols, we provide greater details of the time protocols that have been used in our architecture. Section II presents a brief review of some of the most popular in-vehicle protocols. We discuss the potential of Ethernet based systems, especially the TT-E standard, for advanced in-vehicle communications. In Section III, we present our proposed network architecture, which includes our novel gateway design for achieving interworking and synchronization between the two time-triggered protocols, namely TT-CAN and

Gharavi, Hamid. "Deterministic Intra-Vehicle Communications: Timing and Synchronization." Paper presented at IEEE International Conference on Communications (IEEE ICC), Dublin, IE. June 07, 2020 - June 11, 2020.

TT-E. Finally, Section IV presents the simulation results of the integrated time protocols followed by the conclusion.

II. Time Triggered Intra-Vehicle Communications

Currently there are a number of intra-vehicle communication protocols such as Controller Area Network (CAN) [4], LIN (Local Interconnect Network) [12], FlexRay [2, 3], Media Oriented Systems Transport (MOST) [13] that have been used in today's vehicle. CAN however, is the most widely utilized technology. It is available in different forms, such as low and high-speed CAN, with Flexible Data Rate. Low-speed CAN has a data rate of 40 Kbps to 125 Kbps while the high -speed CAN offers bandwidth from 40 Kbps to 1 Mbps (depending on the length of the cable). In addition, the high-speed Time-triggered CAN (TT-CAN) offers bandwidth from 40 Kbps to 1 Mbps (depending on the length of the cable) [1]. The none-IP compliant TT-CAN has been extensively deployed in today's automobile. Bear in mind that time triggered protocols aim to provide reliable distributed computing and networking for in-vehicle communication systems. In a time-triggered system, the activities are initiated periodically to ensure a high level of determinism. However, the increasing number of ECUs, smart sensors, and other data-generating devices in modern vehicles, would require a much greater bandwidth that the TT-CAN cannot provide.

To meet the demand for greater bandwidth, FlexRay was introduced, which can support data rates of up to 10 Mbps. Some of the key characteristics of the FlexRay protocol are; synchronous and asynchronous frame transfer, guaranteed frame jitter and latency during synchronous transfer, single as well as multi-master clock synchronization, prioritization of frames during asynchronous transfer, error detection and signaling, time synchronization across multiple networks, and scalable fault tolerance [2].

TT-Ethernet (TT-E) is another standard, which is designed to expand traditional Ethernet with services to fulfill the requirements of deterministic, time-critical, and safety-related applications. TT-E is based on the AS6802 standard [11], which has been developed for integrated systems and safety-related applications primarily in aerospace, industrial controls, and automotive applications. It is a Layer 2 Quality-of-Service (QoS) enhancement, which is designed to expand traditional Ethernet-based networks with services to fulfill the requirements of deterministic and time-critical traffic, as well as other types of non-critical traffic. This is achieved by providing different traffic classes in parallel, namely Time-Triggered (TT), Rate-Constrained (RC), and Best Effort (BE). For instance, a time-triggered message, which has the highest priority transmission, is scheduled over the network at pre-configured time instances. The generation (triggered!!) time, delay, and precision of time-triggered messages are pre-set and guaranteed. The rate-constrained messages have less strict determinism and real-time requirements. Rate-constrained messages have a predefined bandwidth with temporal deviations of well-defined bounds. The best-effort messages are similar to traditional ethernet traffic where no delay is guaranteed. These messages use the residual bandwidth of the network and have the lowest priority. TT-E, due to its important features such as low latency, reliability, and higher bandwidth, has been selected as the backbone network in our proposed architecture.

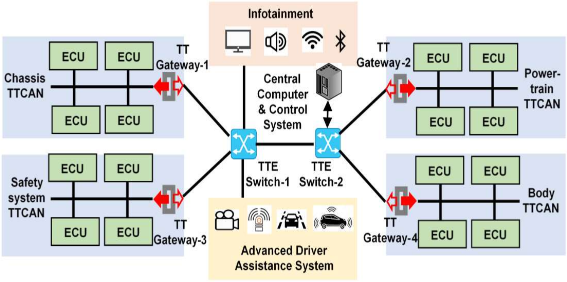


Fig. 1. A simplified in-vehicle time triggered communication system.

III. Proposed Network Architecture

Today's vehicles are becoming more complex than ever mainly due to more advanced requirements for safety regulations, luxury conveniences, computer-based diagnostics, and a vast array of power accessories supported by each car. These requirements have raised a need for more Electronic Control Units (ECUs) in cars. These ECUs not only make monitoring easier, but also report back to the driver if something is wrong. An ECU is a computer module that communicates with the sensors deployed in a car and controls various electrical functions. There are several ECUs in modern cars that control different operations ranging from monitoring engine performance to controlling electronic accessories. Some common ECUs include powertrain control module, chassis control module, advanced driver assistance system, safety system, infotainment module, and comfort module. These ECUs exchange information with each other using various in-vehicle networking technologies. Amongst the most critical in-vehicle components are the advanced driver assistance system and its extension to all aspects of the drive towards "self-driving" cars. The advanced driver-assistance systems (ADAS) mainly focus on enhancing the driving experience by offering technologies to avoid accidents and collisions. These technologies help in detecting potential problems and alerting the driver. The ADAS relies on various types of cameras (e.g., right, left, front, and rear cameras), LADAR, and lane departure warning (LDW)/traffic signal recognition (TSR) systems. The cameras allow the driver to easily detect approaching pedestrians, vehicles, or cyclists with the help of a screen. For instance, the LDW/ TSR mechanisms are used to alert the driver when a vehicle begins to move out of its lane without any prior signal indication. These mechanisms aim at minimizing accidents by relying heavily on the ADAS to guarantee safety of vehicles and passengers and to perform corrective actions such as returning the vehicle to its lane or emergency braking, especially in the case of Self-Driving Automation.

With the growing number of ECUs and further expansion of ADAS for self-driving, the major challenge is how to provide enough bandwidth, as well as ensuring reliability of the in-vehicle communication network. TT-E not only can provide a huge bandwidth, but also offers deterministic communications for safety related sensors and ECUs. The main drawback of the TT-E is the deployment cost (e.g., switches and end-system) compared with the bus-based TT-CAN technology.

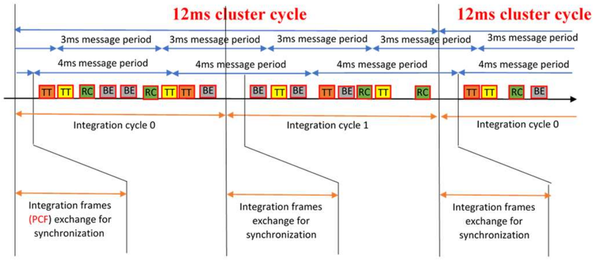


Fig. 2. An example of message periods, cluster cycle and integration cycles.

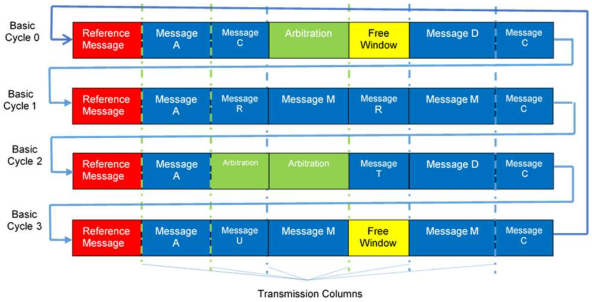
Furthermore, many traditional ECUs have already been designed for deployments in the TT-CAN environments. Therefore, as a comprise solution between cost and capacity, in this paper we a present a network architecture, which uses TT-E as the backbone network with multiple TT-CANs as sub-networks. Fig. 1 shows a simple example of the proposed architecture showing interworking between the two time-triggered protocols: TT-E and TT-CAN, via the design of a suitable gateway that can achieve interoperability between two entirely different time-triggered protocols.

A. *TT-Gateway Design*

The main functionality of the gateway is to allow interworking between two distinctly different time triggered protocols. This would require developing strategies to solve incompatibilities in terms of physical infrastructure, timing, and synchronization. To explore this further, we first provide more detailed information about the timing protocols of TT-E and TT-CAN.

For TT-E synchronization, the AS6802 protocol uses dedicated messages called protocol control frames (PCF) to establish and maintain system-wide clock synchronization amongst all the nodes at regular intervals (e.g., 1 millisecond), which is called integration cycles. The Synchronization entities consist of Compression Master (CM) and Synchronization Master (SM) or synchronization clients (SC), which are selected based on the system architecture. Generally, switches are configured as CM and end-systems as SM. Switches and end-systems that are not configured as synchronization master or compression master will be configured as the synchronization client.

The synchronization is achieved in two steps during normal operation mode (as shown in Fig. 11 of [11]). In the first step the synchronization masters send a PCF message (i.e., a short Ethernet frame: 64 bytes) to the compression masters (CM). The CMs then compute the average value using the relative received times of these PCFs. In the second step, the CMs then send out a new PCF to the SMs and SCs. More specifically, a TT-E switch, as the Compression Master (CM), generates a global time based on the PCF frames received from the Synchronization Masters (SMs) and distributes it to all Synchronization Masters and Synchronization Clients (SCs). In the two-step synchronization, the PCF frame (i.e., exchanges between SM and CM) is known as the integration frame, whereas the synchronization procedure for dispatching these integration frames (sending and compressing) within a configurable period is defined as the integration cycle. Fig. 2 shows the relationship between the integration cycle and the cluster cycle, where the overall cluster cycle in TT-Ethernet consists of multiple integration cycles, ranging from 0 to max_integration_cycle-1 [11]. On the other hand, the cluster cycle comprises the least common multiple of message periods, as depicted in Fig. 2.



Basic Cycle 0	Reference Message	Message A	Message C	Arbitration	Free Window	Message D	Message C
Basic Cycle 1	Reference Message	Message A	Message R	Message M	Message R	Message M	Message C
Basic Cycle 2	Reference Message	Message A	Arbitration	Arbitration	Message T	Message D	Message C
Basic Cycle 3	Reference Message	Message A	Message U	Message M	Free Window	Message M	Message C

Fig. 3. TTCAN system matrix.

Unlike TT-E, TT-CAN is a serial bus network and its messages are broadcast on the bus (instead of using sender and receiver's addresses) where they can be picked up by interested receivers according to the message's unique identifier. As a TDMA based scheme, TT-CAN divides the timeline between two consecutive reference messages, called the basic cycle, into time slots (windows). As shown in Fig. 3, a system matrix can be constructed by grouping a number of basic cycles. A basic cycle comprises several time windows of different sizes where a time master can send a reference message to every node to achieve synchronization. A time window may be an exclusive window, an arbitration window, or a free window. Exclusive windows are mainly considered for periodic messages without having to compete for network access (i.e., deterministic). An arbitrating window is for event triggered messages and a free time window is reserved for future extensions. A TT-CAN node, however, does not need to know the whole system matrix (only the information in each message) and based on the reference message, nodes can update their local time slots for transmission and reception of their data messages.

There are two possible timing levels in TT-CAN. In the first level (level-1) a time-triggered operation is carried out based on the reference message from a time master while an independent clock runs in each node. On the other hand, level-2, with the support of global time and a continuous drift correction in each node, provides higher synchronization quality. In our model, level 2 timing is considered where the TT-CAN section of the gateway operates as the time master in the TT-CAN network. Under level-2 timing, each node uses a cyclically incrementing counter as its Local-Time, which is decided by the Network Time Unit (NTU). The local time contains at least 19 bits where the three least significant bits represent a fraction of the NTU and is incremented in the unit of $NTU/2^n$. The length of the NTU is configurable and generated locally based on the local Time Unit Ratio (TUR) where $NTU = TUR \cdot t_{sys}$ and t_{sys} is the local clock period. It should be noted that TUR is a non-integer value and can be adjusted accordingly for continuous drift correction [1]. More specifically, TUR can be adjusted based on the difference between local time and the time transmitted by the gateway (i.e., time master).

As shown in Fig. 3, at the beginning of each basic cycle a time master, which is the gateway in our proposed architecture, sends a reference message to every node within its sub-network and start its Cycle_Time (CT). Based on this message, nodes can update their local time slots for the transmission and reception of their messages. When the CT reaches a predefined value: T_{cycle} (the length of the basic cycle), the time master resets its CT, starts a new basic cycle by sending a new reference message to the TT-CAN sub-network. On the other side, when receiving a reference message every node starts its new basic cycle and resets its CT. The value of the local time is saved as the Sync_Mark at the sample point of the Start-Of-Frame (SOF) bit of each message. The Sync_Mark of a reference message is defined as the Ref_Mark.

Fig. 4 shows the level-2 timing process for obtaining the cycle time and global time. As shown, the difference between a node's local time and Ref_Mark is the Cycle_Time, which is reset at the beginning of each basic cycle as soon as Ref_Mark is captured [14]. It should be noted that Global_Time is employed only in level-2 as the reference for the synchronization and calibration of all the local times to the time master's clock, hence providing more fine-grained synchronization. This is generated by the time master and transmitted in the reference message as Master_Ref_Mark to all nodes. The TT-CAN nodes derive their Global_Time by summing their Local_Time and their Local Offset (see Fig. 4). Local_Offset is the difference between the Ref_Mark in Local_Time and the Master_Ref_Mark in Global_Time, caused by the difference between local nodes' NTU and the time master's NTU. In the time master, the Local_Offset is zero. By comparing the differences between two consecutive Master_Ref_Marks (measured in time master's global NTUs and received in reference messages) and two consecutive Ref_Marks (measured in local NTUs), local nodes can derive the clock speed difference and compensate the drift by updating their TURs as follows: $TUR = df \cdot TUR_{previous}$, where drift factor $df = \frac{Ref\_Marks - Ref\_Marks_{previous}}{Master\_Ref\_Marks - Master\_Ref\_Marks_{previous}}$. Local nodes will then update local NTUs after this drift compensation (see Fig. 4) and calibrate local time base to the time master's time base.

B. *Integrated TT-E and TT-CAN*

Figure 5 shows an example of the interworking between TT-E and TT-CAN protocols via a TT-gateway. The function of the gateway is to perform timing and synchronization between the two protocols. Specifically, the TT-E synchronization master in the TT-gateway synchronizes with the TT-E switch (compression master) periodically. Meanwhile the TT-CAN time master in the TT-gateway updates its clock to the local TT-E clock at the beginning of each basic cycle to achieve synchronization with the compression master (as seen in Fig. 5). Under these conditions, a TT-E switch, as the Compression Master (CM), generates a global time that is based on the PCF frames received from SMs, which is then distributed to all SMs and SCs. The synchronization process within the integration cycle operates inside a configurable period where TT-E clocks in TT-gateways can be updated. On the other hand, before sending the Reference (REF) message to the local TT-CAN network, the TT-CAN time master in the TT-gateway updates its clock to the local TT-E clock at the beginning of every basic cycle. The period in which the TT-CAN can be updated depends on the total amount of aggregated TT-CAN data that can be generated by multiple ECUs in the TT-CAN subnetwork. Therefore, due to the much larger bandwidth of the TT-E backbone network, such a time interval should be much larger than the TT-E integration cycle. As an example, in our simulation, the TT-CAN is updated every 0.024576 seconds (configurable) with respect to the TT-E integration cycle of 0.003.

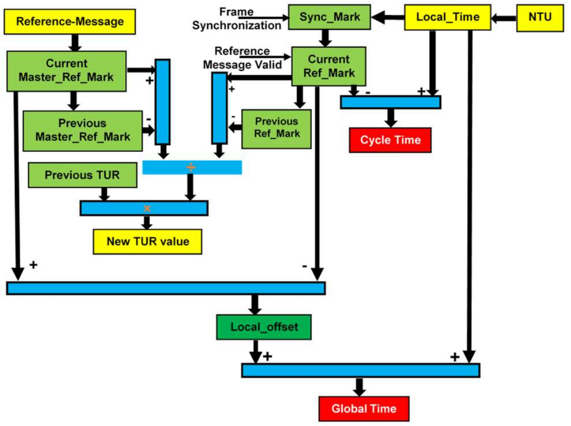


Fig. 4. Level-2 timing process: Cycle_Time, Global_Time, Local_offset and New TUR.

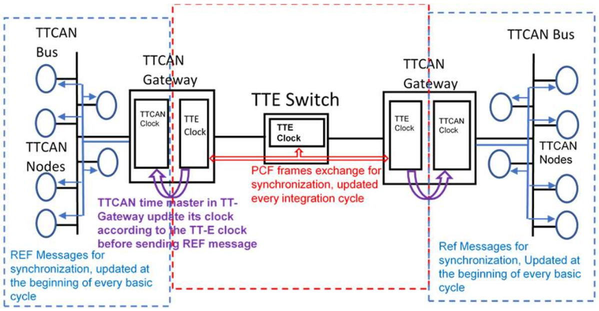


Fig. 5. An example of interworking between TT-E and TT-CAN protocols via a TT-Gateway.

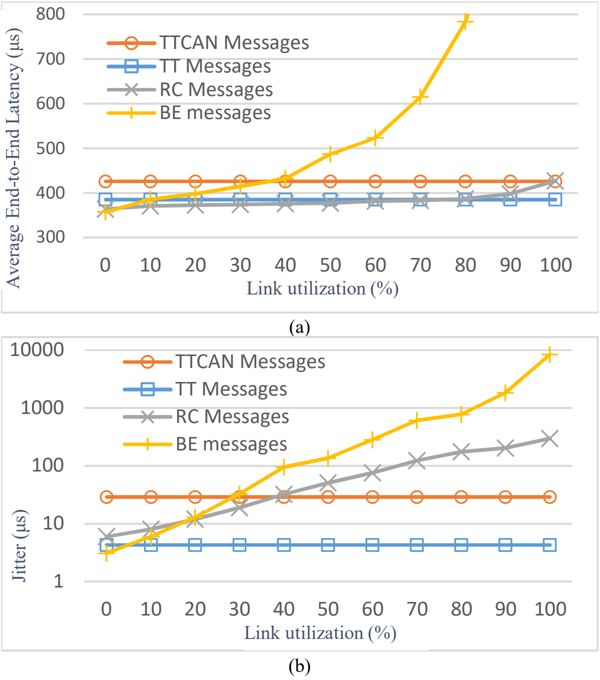
Without an external clock, the TT-CAN time master can keep its TUR and NTU constant. However, in our system, the time master (the TT-CAN clock in the TTCAN gateway) will update itself with the TT-E clock in order to achieve full synchronization in the proposed integrated TT-E and TT-CAN network. At the end of the basic cycle, the TT-CAN clock in the TT-gateway first updates its Local_Time by incorporating the difference between itself and the TT-E clock; T_{gap}, where

$$\text{Local\_Time} = \text{Local\_Time}_{previous} + T_{gap}.$$

The new Local_Time is saved as the Master_Ref_Mark. It then updates its

$$df = \frac{Local\_Time_{previous} - Master\_Ref\_Marks_{previous}}{Maste\ \ \_Ref\_Marks - Maste\ \ \_Ref\_Marks_{previous}},$$
$$TUR = df \cdot TUR_{previous}$$

Gharavi, Hamid. "Deterministic Intra-Vehicle Communications: Timing and Synchronization." Paper presented at IEEE International Conference on Communications (IEEE ICC), Dublin, IE. June 07, 2020 - June 11, 2020.



(a)

(b)

Fig. 6. The influence of link utilization on (a): average end-to-end latency performance and (b): the jitter performance of all traffic.

and NTU, calibrating their time base to that of the TT-E backbone network. Furthermore, it sends a new reference message to TT-CAN sub-network and resets its Cycle_Time to T_{gap} (not zero as before). This means that the TT-CAN clock in TT-gateway will be re-synchronized with the TT-E clock after $T_{cycle} - T_{gap}$ seconds.

When receiving reference messages, the local node in the TT-CAN sub-network will first update its df as

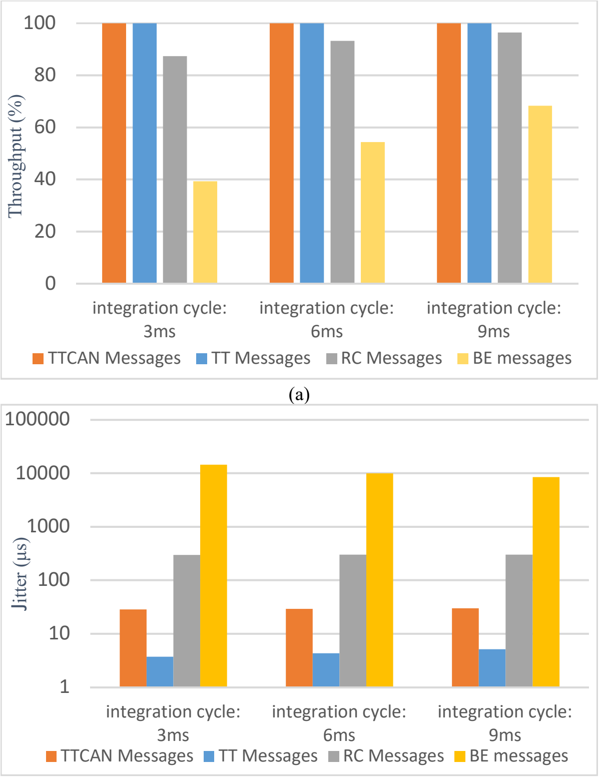
$$df = \frac{Local\_Time - Ref\_Marks_{previous}}{Master\_Ref\_Mar \qquad gap - Master\_Ref\_Marks_{previous}}.$$

It then updates TUR and NTU. Note that the time master has updated its Local_Time and Master_Ref_Marks by incorporating the T_{gap} between the TT-CAN sub-network and the TT-E backbone network. The local node then updates its Local time to Master_Ref_Marks + τ (τ is the transmission delay between the local node and the time master) and saves it as Ref_Mark, to achieve synchronization with the time master.

IV. Simulation

In this section, we evaluate integrated time triggered communication in the system topology of a simplified in-vehicle communication network shown as in Fig. 1, where video, audio, control, and inter-gateway traffic are categorized as BE traffic, RC traffic and TT traffic for investigation.

As shown in Fig. 1, the integrated in-vehicle communication network consists of two TT-E switches and 4 TT-Gateways producing a TT-E backbone network. Each TT-Gateway interconnects a TT-CAN sub-network with the TT-E backbone network. High-speed devices, such as ADAS and infotainment modules, are directly connected to one of the TT-E switches, sending TT traffic, RC traffic, and BE traffic to the control system through TT-E links. All TT-E links are bidirectional 100 Mbps links. On the other hand, low-speed devices, such as ECUs are connected to TT-CAN sub-networks with a 1 Mbps bandwidth. In our simulation, TT-CAN messages between ECUs are first sent to the local TT-Gateway and converted to TT messages, which will be sent to destination TT-Gateways via TT-E switches. A destination TT-Gateway first converts them back to TT-CAN messages before forwarding them to the destination ECUs or the central control unit.



(a)

(b)

Fig. 7. The impact of the integration cycle on (a): throughput performance and (b): jitter performance.

For simplicity, the payload of the TTCAN traffic is set at 8 bytes, the payload of the TT traffic is 50 bytes, the payload of the RC traffic is 100 bytes, while the payload of the BE traffic is 500 bytes. The drifts of TT-E clocks and TT-CAN clocks are configured as 200 parts per million (ppm). The propagation delay of the TT-E backbone network is 100 ns per link, while the propagation delay of the TTCAN bus is set to 100 ns (5 ns per meter propagation delay and a maximum cable length of 40 meters on the CAN bus with 1 Mbps bandwidth [4]).

In a TT-E network, virtual links are used as logical connections to route traffic from a sender to one or more receivers and are pre-configured during the scheduling process. TT-E switches use pre-defined forwarding tables to concatenate virtual links and create tree structures with one sender as root and multiple receivers as leaf nodes [15]. Fig. 6 depict the average end-to-end latency and jitter performance of the in-vehicle time triggered communication network under the influence of different link utilizations caused by varying amounts of traffic. The TTCAN traffic, as a special and small part of TT Traffic in the TT-E backbone network, is measured with TT traffic but listed separately. Due to synchronization protocols and the link reservation mechanism in both TT-E backbone networks and TT-CAN sub-networks, TTCAN traffic and TT traffic are able to achieve a constant average end-to-end latency with very low jitter, despite the increasing link utilization. Their low jitters are mainly due to the clocks' drift.

Since TTCAN traffic experiences extra links with a lower timing resolution in TT-CAN sub-networks, it has a higher jitter (28.97 μs) and latency (426 μs) compared with TT traffic (i.e., jitter of 4.3 μs and latency of 385 μs). The jitter caused by the TT-CAN clocks' drift is higher than the jitter caused by the TT-E clocks' drift. The basic cycle (e.g. 0.024576 s) in TT-CAN is much longer than the integration cycle (e.g. 0.024576 s) in TT-E, resulting in more drift in TT-CAN clocks than in TT-E clocks before clock correction. The latency and jitter performance of the RC traffic degrades because of increasing link utilization. However, it is limited to an ensured bounded latency. For BE traffic, its latency and jitter performance degrade significantly (with increasing traffic) when the link overload is increased. The fully utilized link results in high collision possibility and unstable communication.

Our investigation in Fig. 7 is based on worst-case scenarios. In these experiments, we allocate 1.6Mbps, 18 Mb/s, and 30 Mb/s bandwidth to handle TTCAN, TT, and RC traffic, respectively. In Fig. 7, we evaluate the influence of the integration cycle on the performance of all traffic in a worst-case scenario. When reducing the period of the integration cycle, more overhead will be generated because of the PCF frames exchanged between the compression master and synchronization masters. On the other hand, a clock's drift is decreased with a shorter integration cycle and therefore improves the timing precision in the TT-E backbone network. It can be seen from Fig. 7(a) that TTCAN and TT traffic is not affected by the varying integration cycle because they are protected by link reservation. The throughput performance of RC and BE traffic is degraded due to the increased overhead. Fig. 7(b) demonstrates the jitter performance under the influence of the integration cycle. TTCAN, TT and RC traffic achieves a slight improvement in the jitter performance due to increased time precision when reducing the integration period. In contrast, BE traffic shows a worse jitter performance due to the dominant factor of increased overhead and packet loss rate.

V. Conclusion

In this article, we present a time triggered communication network architecture that integrates two time-triggered protocols: namely TT-Ethernet and TT-CAN. Specifically, high speed TT-E is employed as the backbone network to comply with the high bandwidth requirement, while lower bandwidth TT-CAN is used as sub-networks for low speed ECUs with consideration of cost-effectiveness. Thanks to the synchronization and link reservation mechanism in both TT-E and TT-CAN, they are capable of supporting constant and low latency traffic with very low jitter. This guarantees reliability in-vehicle communication. Our investigation into fully overloaded links and worst-case scenarios further demonstrates the reliability and robustness of the integrated time-triggered communication network architecture.

Acknowledgment

The authors would like to thank Dr. Ejaz Ahmed, a former NIST associate, for his involvement in the initial stages of this work.

References

[1] Road vehicles – Controller area network (CAN) – Part 4: Time triggered communication; ISO 11898-4:2004.

[2] FlexRay consortium, "Protocol Specification", Specification 2.1, Stuttgart, Germany, Dec 2005.

[3] R. Shaw and B. Jackman, "An introduction to FlexRay as an industrial network," 2008 IEEE International Symposium on Industrial Electronics, Cambridge, 2008, pp. 1849-1854.

[4] Road vehicles – Interchange of digital information – Part 1: Controller area network data link layer and medium access control; ISO 11898.

[5] M. Khader and S. Cherian, "An Introduction to Automotive LIDAR - Texas Instruments," http://www.ti.com/lit/wp/slyy150/slyy150.pdf.

[6] H. Gharavi and S. Gao, "3-D Motion Estimation Using Range Data," in IEEE Transactions on Intelligent Transportation Systems, vol. 8, no. 1, pp. 133-143, March 2007.

[7] S. Tuohy, M. Glavin, C. Hughes, E. Jones, M. Trivedi and L. Kilmartin, "Intra-Vehicle Networks: A Review," in IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 534-545, April 2015.

[8] D. Thiele, J. Schlatow, P. Axer and R. Ernst, "Formal timing analysis of CAN-to-Ethernet gateway strategies in automotive networks." Real-Time Syst. January 2016, Volume 52, Issue 1, pp 88–112.

[9] H. Gharavi ; K. V. Prasad ; P. Ioannou , "Scanning Advanced Automobile Technology," Proceedings of the IEEE , Volume: 95 , Issue: 2 , Feb. 2007

[10] Ixia. "Automotive Ethernet primer whitepaper." Internet: www.ixiacom.com, May 2014.

[11] SAE AS6802 Standard, www.sae.org.

[12] Road vehicles – Local Interconnect Network (LIN) -- Part 1: General Information and use case definition; ISO 17987-1:2016.

[13] Grzemba, Andreas (2011). MOST: The Automotive Multimedia Network; from Most25 to Most150. Poing: Franzis. ISBN 978-3-645-65061-8.

[14] F. Hartwich, B. Müller, T. Führer, R. Hugel and R. Bosch, "Timing in the ITCAN Network", Robert Bosch GmbH; Proceedings 8th International CAN Conference; 2002; Las Vegas.

[15] D. Tamas, P. Pop, W. Steiner, "Synthesis of communication schedules for TTEthernet-based mixed-criticality systems", *Proc. 8th IEEE/ACM/IFIP Int. Conf. Hardware/Software Codes. Syst. Synthesis*, pp. 473-482, 2012.

The More the Merrier: Adding Hidden Measurements for Anomaly Detection and Mitigation in Industrial Control Systems

Jairo Giraldo
Electrical Engineering Department
University of Utah
jairo.giraldo@utah.edu

David Urbina
Computer Science Department
University of Texas at Dallas
david.urbina@utdallas.edu

CheeYee Tang
Engineering Laboratory
National Institute of Standards and Technology
cheeyee.tang@nist.gov

Alvaro A. Cardenas
Computer Science and Engineering Department
University of California, Santa Cruz
alvaro.cardenas@ucsc.edu

ABSTRACT

Industrial Control Systems (ICS) collect information from a variety of sensors throughout the process, and then use that information to control some physical components. Control engineers usually have to pick which measurements they are going to use and then they purchase sensors to take these measurements; however, in most cases they only need a small subset of all possible measurements that can be used. Economic and efficiency reasons motivate engineers to use only a small number of sensors for controlling a system; however, as attacks against industrial systems continue to increase, we need to study a systematic way to add sensors to the system to identify potentially malicious attacks. We propose the addition of **hidden sensor measurements** to a system to improve its security. Hidden sensor measurements are by our definition measurements that were not considered in the original design of the system, and are not used for any operational reason. We only add them to improve the security of the system and using them in anomaly detection and mitigation. We show the addition of these new, independent, but correlated measurements to the system makes it harder for adversaries to launch false-data injection stealthy attacks and even if they do, it is possible to limit the impact caused by those attacks. When an attack is detected we replace the compromised sensor measurements with estimated ones from the new sensors improving the risky open-loop simulations proposed by previous work.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

KEYWORDS

ICS, Security, CPS

ACM Reference format:
Jairo Giraldo, David Urbina, CheeYee Tang, and Alvaro A. Cardenas. 2020. The More the Merrier: Adding Hidden Measurements for Anomaly Detection and Mitigation in Industrial Control Systems. In *Proceedings of Hot Topics in the Science of Security, Lawrence, KS, April 7–8, 2020 (HoTSoS '20),* 10 pages.
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Attacks against the integrity of cyber-physical systems are a growing concern. An attacker that falsifies the data that sensors are reporting or falsifies the actions that actuators are supposed to execute, can drive the system to unsafe states, causing potential operational, economic, and safety problems. An attacker can compromise a subset of sensors and send false information to the control system. The attacks do not even have to be because of software vulnerabilities; new *transduction attacks* [10] allow the attacker to change sensor signals without compromising any device. Sensors are transducers that translate a physical signal into an electrical one, but these sensors sometimes have couplings between the property they want to measure and the analog signal that can be manipulated by the attacker. For example, sound waves can affect accelerometers and make them report incorrect movement values [20], and radio waves can trick pacemakers into disabling pacing shocks [16].

To detect these attacks we can use our understanding of the physical evolution of the system, to see if the measurements from sensors match our predictions. There is an active community working on this type of Physics-Based Attack Detection systems (PBAD) [11]. PBAD has been explored in water control systems [1, 13], state estimation in the power grid [9, 17], chemical processes [2, 4], autonomous vehicles [6], and a variety of other cyber-physical systems [11].

All these models assume that the sensors we use for attack-detection are the same that are already present for the control algorithm. Furthermore, attack-mitigation proposals like Cardenas et al. [4] remove the sensor under attack and estimate the missing quantity with the remaining sensor measurements (they try to operate the system with less information, given the removal of this measurement). However, they do not take into account that we can gather new measurements, usually from different stages of a cyber-physical system which are correlated with each other.

We call these new measurements **hidden sensor measurements** because they are hidden from the operation of the system under normal conditions. Furthermore, because hidden measurements are not used in regular operations, an attacker that performs reconnaissance of the industrial network will not see them (hidden sensors may remain silent and will only start reporting values based

Giraldo, Jairo; Tang, CheeYee; Urbina, David; Cardenas, Alvaro. "The More the Merrier: Adding Hidden Measurements for Anomaly Detection and Mitigation in Industrial Control Systems." Paper presented at Hot Topics in the Science of Security (HotSoS) 2020, Lawrence, KS, US. April 07, 2020 - April 08, 2020.

on a request by the intrusion mitigation algorithm or by the intrusion detection system based on some indicators of compromise).

For example, in the classical Tennessee-Eastman (TE) chemical process benchmark [19] (which has been used extensively in cyber-security [2, 4, 8]), the "separator cooling water outlet temperature" is not measured nor used for any purpose; however we found that this variable is highly correlated with the "product separator temperature," a measurement that is critical for safe control of the system. If an attacker falsifies or takes down the "product separator temperature," we can use the "separator cooling water outlet temperature" (with an appropriate estimation algorithm) to derive a good estimate of the attacked-sensor. Similarly the "pressure in the stripper" is a measurement that is not used at all in any control loop; however this measurement is highly correlated with the "pressure in the reactor," and therefore we can use the pressure in the stripper for security purposes.

Our contributions include:

- We introduce the concept of **Hidden Measurements**; i.e., new measurements that are not used during the normal operation of the system, but only for security purposes.
- Using new hidden sensors, we propose an anomaly detection architecture that uses the correlations between *operational* sensor measurements and *hidden* sensor measurements. Our formulation is generic and can be applied in a wide number of cyber-physical systems.
- We introduce a mitigation strategy that uses hidden sensors to respond to an attack. In particular, when an attack is detected, we generate approximate sensor signals based on redundant sensors using autoregression models. For instance, in the TE benchmark, if the reactor pressure value is compromised, we can use our added stripper pressure sensor or added separator pressure sensor (which are not used in any control loop [19]) to estimate the reactor pressure value.
- We implement our attacks and defenses in a Hardware-in-the-Loop testbed that uses a TE process simulation and is controlled with an industrial Programmable Logic Controller (PLC). In particular, we show how to launch man-in-the-middle attacks against an Open Platform Communications (OPC) server to coordinate the communication between the central controller and the field devices. The attack is able to intercept Ethernet/IP packets and falsify sensor/actuator information.

2 PROBLEM FORMULATION

In this section, we present a general mathematical model of an industrial control system (ICS) with an anomaly detection scheme. The dynamics of an ICS can be modeled using differential equations as

$$\begin{aligned}\dot{\boldsymbol{x}}(t) &= F(\boldsymbol{x}(t), \boldsymbol{u}(t)),\\ \boldsymbol{y}(t) &= H(\boldsymbol{x}(t), \boldsymbol{u}(t))\end{aligned} \tag{1}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$ corresponds to the vector with the states of the process (e.g., temperature, pressure, and water level in a chemical reaction), $\boldsymbol{u}(t) \in \mathbb{R}^m$ describes the control commands, and $\boldsymbol{y}(t) \in \mathbb{R}^p$ are the sensor readings. $F(\cdot)$ and $H(\cdot)$ are nonlinear functions that describe the system behavior.

Due to the complexity of some industrial processes, decentralized control has proven to be a practical control option in industrial plants. Decentralized control has many benefits, including easy implementation, maintenance, tuning, and robust behavior. For this reason, we assume that there are m decentralized control loops, and each one has been designed and tuned to guarantee specific performance conditions, as depicted in Figure 1; this architecture is usually implemented with Programmable Logic Controllers (PLCs), each of them controlling a subset of a larger infrastructure. In addition, we assume that there has been an adequate control configuration or input-output pairing, such that for each input u_i there is a suitable output y_i [15, 18]. We then define the controller as follows

$$u_i(t) = \mathcal{K}_i(y_i(t)), \tag{2}$$

where $\mathcal{K}_i(\cdot)$ is the function that takes sensor readings and generates control commands.

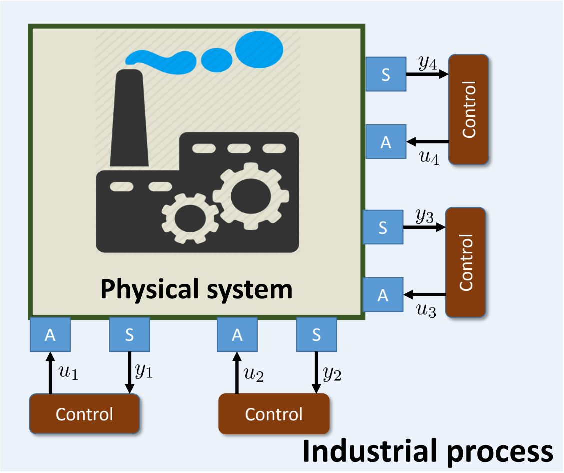


Figure 1: Decentralized control loops in an industrial control system.

Let us assume that the sensor readings and control commands can be sampled each τ seconds such that we have $y_i(k)$ and $u_i(k)$, for $k \in \mathbb{Z}_+$ the k^{th} sampling instant. Using the historical data of the sensors and control actions, we can define an estimation function of the form

$$\begin{aligned}\hat{\boldsymbol{x}}(k+1) &= \hat{F}(\boldsymbol{y}(k), \boldsymbol{y}(k-1), \ldots, \boldsymbol{y}(k-T), \boldsymbol{u}(k), \boldsymbol{u}(k-1), \ldots, \boldsymbol{u}(k-q))\\ \hat{\boldsymbol{y}}(k) &= \hat{H}(\hat{\boldsymbol{x}}(k), \boldsymbol{u}(k))\end{aligned} \tag{3}$$

where T and q are the number of historical values we consider for the sensors and the actuators (respectively). Notice that Kalman filters, ARMA models, and neural networks can be described using Equation (3).

2.1 Detection Mechanism

Physics-based anomaly detection strategies compare current sensor readings with an estimation or prediction of the behavior of the system to detect cyber-attacks [11]. This prediction can be computed using approximated models as described in equation (3). Thus, for each sensor measurement we define the residual $r_i(k) = y_i(k) - \hat{y}_i(k)$ as the difference between the sensor reading and its corresponding prediction. An anomaly detection metric $\mathcal{D}(r(k))$ quantifies how different the historical behavior the sensor reading is from the predicted one. For instance, the χ^2-detection computes the normalized summation of all the residuals, $\mathcal{D}(r(k)) = r(k)^\top \Sigma r(k)$, for Σ the inverse of the covariance matrix. More sophisticated mechanisms accumulate the residuals in order to keep historical track of possible persistent attacks such as the CUSUM algorithm [4]. In this work, we focus our attention on the distributed bad-data detection (DBDD) mechanism defined by $\mathcal{D}_i(r_i(k)) = |r_i(k)|$. DBDD provides a detection metric for each sensor, which is useful in the context of decentralized ICS.

2.2 Attacker Model

We consider a powerful adversary that gains access to a subset of sensors and/or actuators. The adversary knows the detection architecture, the prediction algorithms we use, and is able to generate accurate sensor predictions.

3 ADDING HIDDEN SENSORS

ICS use sensor readings in order to generate control actions; however, most systems possesses physical quantities in the system that are not used, and are not necessarily needed for the safe operation of a system (without considering attacks).

However, if we have a physical model of the system, we can identify variables that although not needed, might be useful for security purposes. Suppose that in our model of the system we have the following observable variables (variables for which a physical sensor can be bought and installed to measure): $\mathcal{I}^y = \{1, \ldots, p\}$. Let $\mathcal{I}^{CL} \subset \mathcal{I}^y$ be the indexes of the variables that are currently being measured and that belong to a control loop and are paired with a control input. Similarly let $\mathcal{I}^{rd} \subset \mathcal{I}^y$ be the indexes of variables that can be physically measured, but that we are not currently measuring in our physical system. Notice that $\mathcal{I}^y = \mathcal{I}^{CL} \cup \mathcal{I}^{rd}$.

According to equation (2), the control command $u_i(k)$ depends on $y_i(k)$, for all $i \in \mathcal{I}^{CL}$. As a consequence, attacks in $y_i(k)$ will affect the control command causing the system to deviate from its operation point. In order to leverage our new hidden sensors we need to find potential sensor signals that are also affected by u_i.

There are several techniques that help to quantify the input-output relationship, such as Relative Gain Array (RGA) and its variations for linear and nonlinear systems [15]; however, they depend on having a very accurate dynamic model of the system, which is difficult for nonlinear industrial control systems. We propose a simple approach that uses historical data of sensor readings and consists on calculating correlation coefficients between each pair (y_i, y_j) for $i \in \mathcal{I}^{CL}$ and $j \in \mathcal{I}^{rd}$. The correlation coefficient determines the degree at which two variables' movements are associated. As a consequence and because of the feedback relationship between y_i, u_i, if the pair (y_i, y_j) is highly correlated, this implies that the control action u_i affects not only y_i but also our potential hidden sensor measurement y_j.

Let $s_i \in \mathbb{R}^T$ be the signal that consists on T readings of sensor i under normal operation. The correlation coefficient is then calculated as follows

$$corr(s_i, s_j) = \frac{cov(s_i, s_j)}{\sqrt{cov(s_i, s_i)cov(s_j, s_j)}}$$

where $cov(s_i, s_j)$ denotes the covariance between s_i and s_j and correlation ranges between $-1 \leq cor(s_i, s_j) \leq 1$.

Next, we will introduce how we can use the correlation coefficient to build multi-variable anomaly detection algorithms that take advantage of hidden sensors.

3.1 Multi-Variable Anomaly Detection

Typically, centralized anomaly-detection mechanisms gather all sensor readings to construct a single prediction model in order to compute residuals and calculate a detection metric (e.g., χ^2); however, for systems with a large amount of sensors and several interconnected processes, these kind of models can be computationally expensive. Taking advantage of the decentralized nature of multiple control loops and redundant sensors, it is possible to decrease the complexity of the prediction models by constructing individual models for each control loop. Each model can be implemented locally at each control loop, and it does not only help to reduce the computational cost but also removes the single point of failure. The estimation of sensor i can be obtained according to

$$\hat{y}_i(k) = \hat{h}_i(y_i(k-1), y_i(k-2), \ldots, y_i(k-T), u_i(k-1), \ldots, u_i(k-q)).$$

The main limitation of this approach lies in the fact that an intelligent adversary can easily design stealthy attacks by only affecting a single sensor. The main idea behind multi-variable detection lies in combining some properties of both approaches, centralized and decentralized by using hidden sensors in order to limit the impact of cyber-attacks.

Our proposed Multi-Variable Detection (MVD) architecture is depicted in Figure 2 and consists on building a single prediction of sensor i based on the history of the control commands, y_i readings, and redundant sensors measurements (to ease notation we refer to redundant sensors as y_j^{rd}). The main difference with a centralized strategy is that the prediction model only depends on highly correlated sensors, instead of all sensors. As a consequence, the complexity of the model is much lower and still guarantees good accuracy.

If an adversary attacks $y_i(k)$, the controller u_i will be affected, which in turn will also affect the redundant sensors. As a consequence, the effects of the attack in y_i and in all y_j^{red} will add up causing an error in the prediction.

Giraldo, Jairo; Tang, CheeYee; Urbina, David; Cardenas, Alvaro. "The More the Merrier: Adding Hidden Measurements for Anomaly Detection and Mitigation in Industrial Control Systems." Paper presented at Hot Topics in the Science of Security (HoTSoS) 2020, Lawrence, KS, US. April 07, 2020 - April 08, 2020.

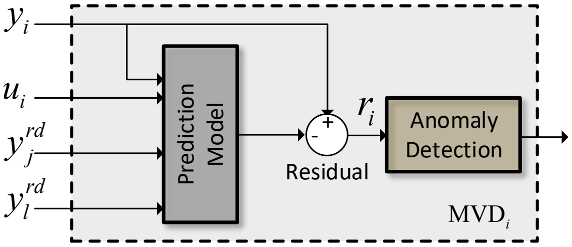


Figure 2: General architecture of the Multi-variable detection block.

3.2 System Reconfiguration for Attack Mitigation

When an attack is successfully detected, it is necessary to remove the compromised sensor readings while maintaining the system operation as close as possible to the nominal operation. In one of our earlier papers [4], we proposed to replace compromised sensor readings with estimated ones obtained from the remaining sensors. In this work, we propose a different approach that makes use of new hidden sensors. Let y_i denote sensor i and y_j^i denote one of the hidden and correlated sensors to i. Therefore, it is possible to find a mapping of $y_j^i \rightarrow \widetilde{y}_i$, where $\widetilde{y}_i$ is an approximation of the sensor reading y_i. This mapping can be computed using historical data and autoregressive models or by knowing the physical relationship between the two sensors (e.g., it is possible to compute pressure from temperature in a gas using the Gay-Lussac's Law.)

Then, when an attack associated to sensor i is detected, we replace the compromised sensor reading $y_i(k)$ with its approximation $\widetilde{y}_i(k)$ to ensure the operation of the system. The higher the correlation, the better the approximation. This is illustrated in Figure 3.

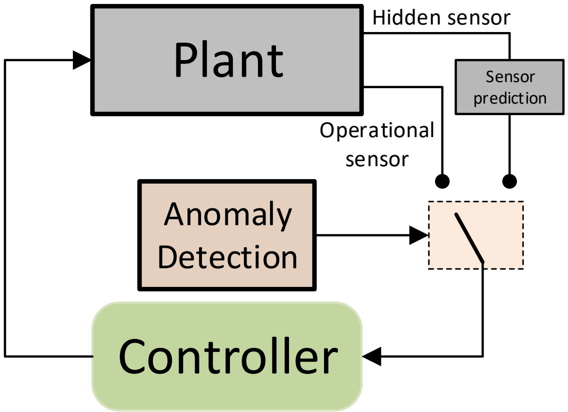


Figure 3: The system is reconfigured to use hidden sensors.

4 TESTBED

4.1 Description

The Tennessee-Eastman (TE) process was first proposed by Down and Vogel [7] and has been extensively used for the evaluation of novel control techniques, due to its complexity and large number of sensors. We were one of the first groups to use this process to study the security of industrial control systems [4, 14].

The process has five major unit operations: the reactor, the product condenser, a vapor-liquid separator, a recycle compressor, and a product stripper. The process produces two products, G and H, from four reactants A, C, D, and E. It has 41 measurements and 12 manipulated variables. The TE is open-loop unstable, which makes it very sensitive to cyber-attacks that affect the control actions.

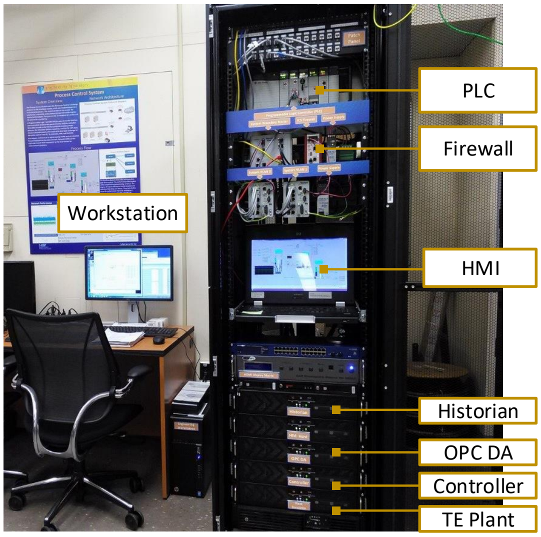


Figure 4: Tennessee-Eastman HIL Testbed.

The NIST Hardware-In-the-Loop (HIL) testbed for the TE process consists of 5 modules running in Microsoft Windows machines:

- simulated plant in C++,
- OLE for Process Technology (OPC) Server,
- distributed proportional-integral-derivative (PID) controller proposed in [19] where 12 control loops keep the states of the plant within operational limits while desired set-points are followed.
- Historian,
- Human-Machine Interface (HMI).

The network architecture of the testbed is illustrated in Figure 5.

The testbed also has an Allan Bradley Programmable Logic Controller (PLC). The C++ plant simulation interacts with the PLC through a Common Industrial Protocol (CIP) communication link. The PLC interacts with the OPC Server through an OPC communication link. The OPC Server communicates with the controller, the

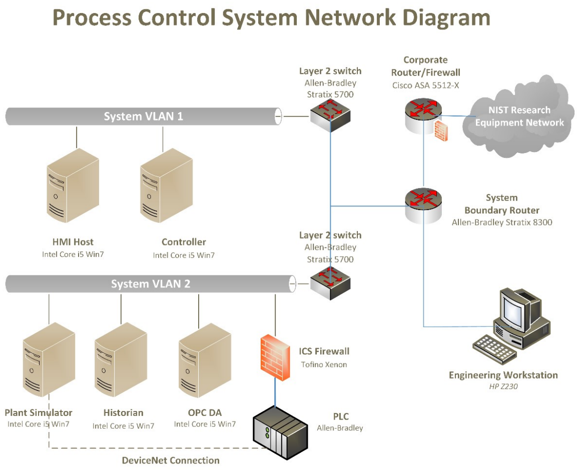


Figure 5: Network architecture of the TE testbed.

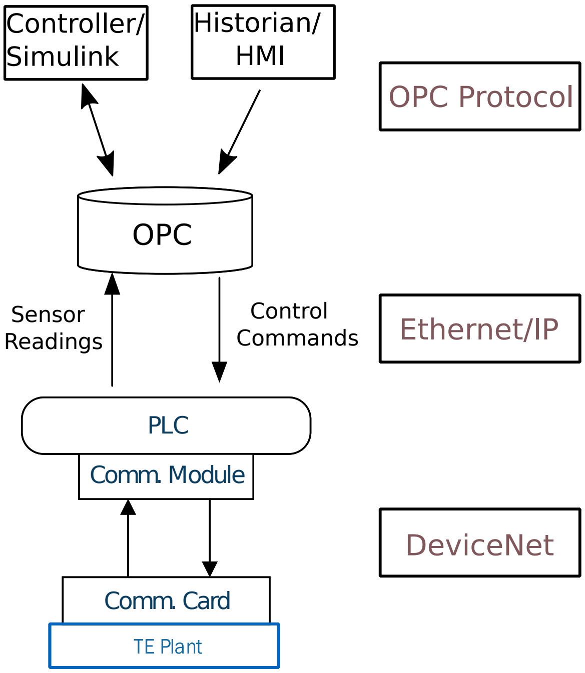
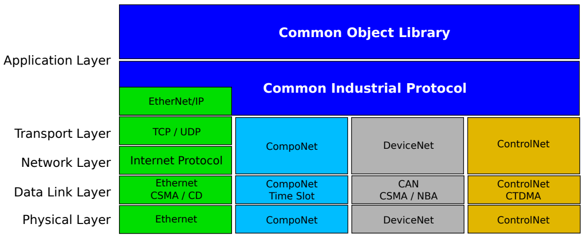


Figure 6: Logical architecture of the TE testbed.

Historian, and the HMI through the TCP/IP network. This logical interaction of components is illustrated in Figure 6.

We believe this testbed represents highly relevant aspects of an ICS. Field communications (Layer 0) are captured by the DeviceNet protocol; industrial network protocols used to connect PLCs and workstations (Layer 1) is represented by the Ethernet/IP industrial protocol, and the widely available OPC server is used to translate among different standards and technologies. The OPC protocol is under particular interest for study as it is one of the industrial protocols that was targeted by the Industroyer malware that attacked Ukraine's power grid in 2016 [5]. OPC was also targeted by the Havex industrial espionage malware [23].

4.2 Common Industrial Protocol



Application Layer	Common Object Library			
	Common Industrial Protocol			
	EtherNet/IP			
Transport Layer	TCP / UDP	CompoNet	DeviceNet	ControlNet
Network Layer	Internet Protocol			
Data Link Layer	Ethernet CSMA / CD	CompoNet Time Slot	CAN CSMA / NBA	ControlNet CTDMA
Physical Layer	Ethernet	CompoNet	DeviceNet	ControlNet

Figure 7: CIP stack and its different physical layers.

The Common Industrial Protocol (CIP) network specification library [3] was originally developed by Rockwell Automation and subsequently standardized and maintained by Open Device Vendors Association (ODVA) and ControlNet International. It aims to fulfill the main three needs of ICS systems: control, configuration, and collection of data. It defines the CIP application layer protocol as a encapsulated object-oriented protocol for transmission of connected (I/O implicit) messages between a data producer and one or more data consumer devices, and unconnected (explicit) messages between two devices in the control network. Transmissions associated with a particular connection are assigned a unique *connection ID*. While being an application layer protocol, CIP is independent of the underlying layers, and requires an encapsulation protocol which allows abstraction from different data link and physical layers. It also includes a Common Object library defining commonly used objects, some of which are specific for a particular encapsulation protocol, and allows for extension and definition of vendor specific objects. The CIP specification library includes the definition of 4 different CIP stacks depending of the physical layer in use (see Figure 7): Ethernet/IP(over IEEE 802.3 Ethernet), CompoNet, DeviceNet, and ControlNet.

4.2.1 Ethernet/IP. The CIP stack introduces the Ethernet/IP protocol [3] for both, SCADA network and fieldbus communications alike. Its specification defines the Common Packet Format (CPF) for the encapsulation of message oriented protocols, such as CIP, Modbus, and vendor proprietary messages. Ethernet/IP CPF can be stacked over UDP or TCP, in both multipoint and point-to-point connection modes. When stacking over UDP, it requires devices to select a maximum of 32 consecutive addresses from the range 239.192.1.0 to 239.192.128.255 (which belongs to the *Organizational Local Scope* [12]).

4.3 False-data Injection in CIP packets

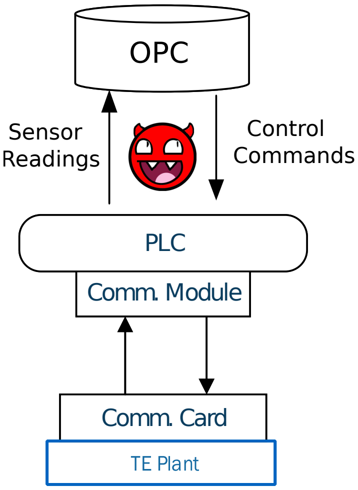


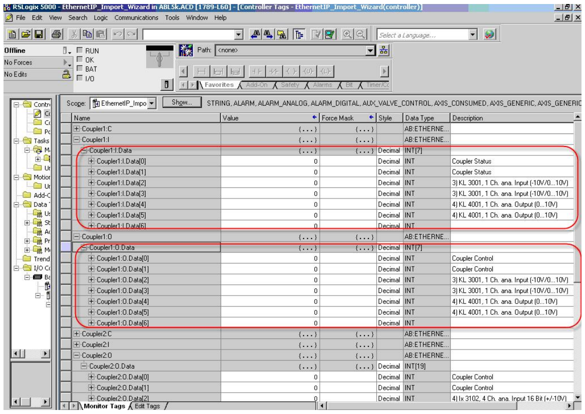
Figure 8: Man-in-the-Middle attack in the TE testbed at NIST.

We established a Man-in-the-Middle attack (MitM) between the C++ plant simulator and the Allan Bradley PLC (Figure 8), by creating a bridge that allows us to capture all the packets coming from the sensors and the controller, modify them and send them back. This communication uses the Common Industrial Protocol (CIP) as industrial communication protocol. Although the CIP specification provides a very comprehensive library of common objects, this communication link implements a vendor-dependent extension over the protocol in order to transfer the 42 sensor measurements constantly by the plant. On the other hand, the 12 actuation commands sent from the PLC to the plant are transmitted with separate and standard CIP write-object messages containing the actuator ID number and the command value.

The CIP communication link implements a client/server mode of communication. Therefore, for the MitM to be successful sniffing sensor measurements simulated controller (PLC and OPC server also) must be online. In other words, the controller must request the sensor measurements for our MitM to be able to sniff the response from the plant.

We leveraged the Allan Bradley visualization tool (Logix5000 Fig. 9) installed in the workstation to program the PLC to understand at a high level the structure of the CIP extension with the 42 measurements. At first glance, it follows an array structure with every sensor measurement encoded using the Floating-Point Arithmetic Standard IEEE 754.

After developing a Scapy parser for the CIP extension, we performed initial false-data injection attacks. From the results of these attacks, we realized that some sensor measurements (such as the temperature) were being replicated in the CIP extension, and only injecting one of the instances was not enough to successfully achieve the attack, as the controller would freeze the sensor measurement to the last value before the attack started.


RSLogix 5000 - EthernetIP_Import_Wizard in ABL5k.ACD [1789-L60] - [Controller Tags - EthernetIP_Import_Wizard(controller)]
File Edit View Search Logic Communications Tools Window Help
Select a Language...
Offline
No Forces
No Edits
RUN
OK
BAT
I/O
Path: <none>
Favorites
Add-On
Safety
Alarms
Bit
Timer/Co
Contr
Tasks
Motior
Add-C
Data
Trend
I/O Co
Monitor Tags
Edit Tags

Scope: EthernetIP_Impo Show... STRING, ALARM, ALARM_ANALOG, ALARM_DIGITAL, AUX_VALVE_CONTROL, AXIS_CONSUMED, AXIS_GENERIC, AXIS_GENERIC

Name	Value	Force Mask	Style	Data Type	Description
Coupler1:C	{...}	{...}		AB:ETHERNE...	
Coupler1:I	{...}	{...}		AB:ETHERNE...	
Coupler1:I.Data	{...}	{...}	Decimal	INT[7]	
Coupler1:I.Data[0]	0		Decimal	INT	Coupler Status
Coupler1:I.Data[1]	0		Decimal	INT	Coupler Status
Coupler1:I.Data[2]	0		Decimal	INT	3) KL 3001, 1 Ch. ana. Input (-10V/0...10V)
Coupler1:I.Data[3]	0		Decimal	INT	3) KL 3001, 1 Ch. ana. Input (-10V/0...10V)
Coupler1:I.Data[4]	0		Decimal	INT	4) KL 4001, 1 Ch. ana. Output (0...10V)
Coupler1:I.Data[5]	0		Decimal	INT	4) KL 4001, 1 Ch. ana. Output (0...10V)
Coupler1:I.Data[6]	0		Decimal	INT	
Coupler1:O	{...}	{...}		AB:ETHERNE...	
Coupler1:O.Data	{...}	{...}	Decimal	INT[7]	
Coupler1:O.Data[0]	0		Decimal	INT	Coupler Control
Coupler1:O.Data[1]	0		Decimal	INT	Coupler Control
Coupler1:O.Data[2]	0		Decimal	INT	3) KL 3001, 1 Ch. ana. Input (-10V/0...10V)
Coupler1:O.Data[3]	0		Decimal	INT	3) KL 3001, 1 Ch. ana. Input (-10V/0...10V)
Coupler1:O.Data[4]	0		Decimal	INT	4) KL 4001, 1 Ch. ana. Output (0...10V)
Coupler1:O.Data[5]	0		Decimal	INT	4) KL 4001, 1 Ch. ana. Output (0...10V)
Coupler1:O.Data[6]	0		Decimal	INT	
Coupler2:C	{...}	{...}		AB:ETHERNE...	
Coupler2:I	{...}	{...}		AB:ETHERNE...	
Coupler2:O	{...}	{...}		AB:ETHERNE...	
Coupler2:O.Data	{...}	{...}	Decimal	INT[19]	
Coupler2:O.Data[0]	0		Decimal	INT	Coupler Control
Coupler2:O.Data[1]	0		Decimal	INT	Coupler Control
Coupler2:O.Data[2]	0		Decimal	INT	4) Ix 3102, 4 Ch. ana. Input 16 Bit (+/-10V)

Figure 9: Tags visualization and modification tool Logix5000.

To identify the replicated sensor measurements in the CIP extension, we performed a packet diffing similar to the process proposed by Urbina et. al. [21]:

1. Using the Allan Bradley visualization tool we forced 5 different sensor values while sniffing a tuple (2 packets per value change) of CIP generated packets (5 tuples).
2. We then performed diffing of packets to calculate a change-coefficient matrix representation: Any byte-change introduced by the first packet of every tuple increments its cell coefficient, while any change introduced by the second packet of the tuple decrements it.
3. When the 5 tuples were diffed we obtained a matrix with each cell containing the coefficient of change for the corresponding byte in the CIP extension.
4. We used the change-coefficient matrix to visualize the heatmap: the higher the coefficient the higher the heat of the cell (byte), and vice versa.

After understanding the replication on the CIP extension, we improved our parser and were able to launch false-data injection attacks on the sensor measurements.

5 EXPERIMENTAL RESULTS

Due to the correlation of the different distributed control loops, attacks that may be stealthy to one control loop might be visible (easily detectable) in other loops. Under these conditions, adversaries would need to attack all distributed loops simultaneously to remain stealthy, or decrease the impact of such attacks in a way that it does not trigger alarms in other parts of the plant.

The TE benchmark has a large amount of variables where hidden sensors can be deployed. For instance, there are 11 control loops, but 42 measurements. Recall that we define two types of measurements: i) operational sensors i.e., measurements used to generate control

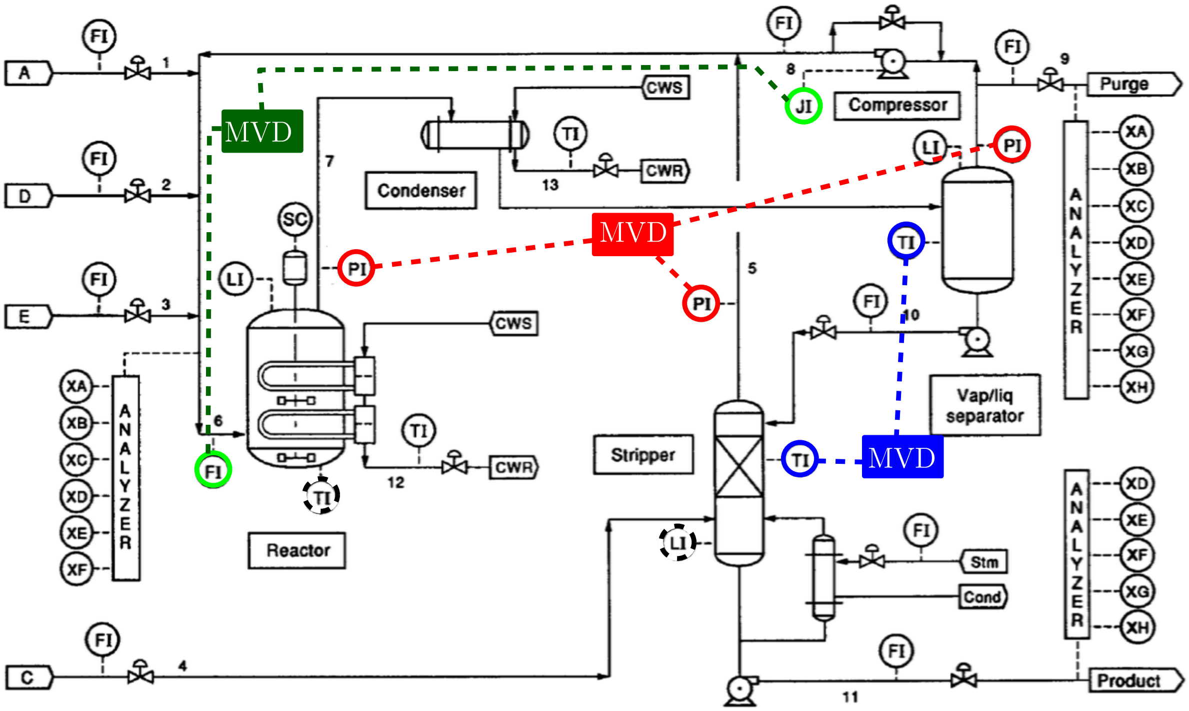


Figure 10: Tennessee Eastman Process with some Multi-variable detection (MVD). Similar colors indicate a high correlation coefficient (i.e., > 0.95). Dashed circles indicate low correlation coefficient (i.e., < 0.01). Each CD receives sensor information and the actuator signal corresponding to at least one of the sensors.

signals, and ii) hidden measurements—that is, measurements that provide information about the system but are not used by the controllers. The correlation among operational and hidden sensors can be exploited to increase the difficulty on deploying stealthy attacks.

Fig. 10 depicts the general architecture of the TE process. Using the correlation coefficient, we are able to identify the hidden sensors with the highest correlation to operational sensors.

As an example, colored circles in Fig. 10 indicate correlation and black dashed circles represent sensors without correlation. Our Multi-Variable Detection (MVD) can be located in such a way it receives information from different PLCs. The information is used to obtain detection statistics for the measurements that are not used. If an attacker wants to remain stealthy, she will have to attack several sensors simultaneously from different PLCs.

Using the identification tool IDENT from Matlab, we are able to obtain individual Hammerstein-Weiner models for several control loops that estimate the input/output relationship. These models combine linear and nonlinear blocks that approximate the sensor measurement behavior for a given control input. Therefore, we are able to generate detection statistics, such as the residuals, by comparing the estimated model with the real system behavior.

As an example, let us consider the reactor pressure (called xmeas 7 in the simulation code), which is a critical parameter of the TE process and is used to control the purge rate, i.e., a valve. We construct a prediction model based on historical data from $xmeas7$ and $xmv6$ using nonlinear ARX models. Figure 11 depicts how a simple bias attack is easily detected by the DBDD strategy. However, an attacker with enough knowledge about the system and the detection strategy is able to design optimal stealthy attacks, as it was proposed in [22], where the adversary replaces the sensor reading by $y_i^a(k) = \hat{y}_i(k) - y_i(k) + \tau_i$, for τ_i the detection threshold. Notice that the residuals under attack become $r_i(k) = |\tau_i|$, such that the detection static is never above the threshold. Now, suppose the adversary forges an attack that remains stealthy for the detector D_7. Figure 12 shows how the attack causes a shut down because the pressure reaches unsafe levels but it is never detected.

Analyzing the correlation coefficients of reactor pressure with measurements that were not being used, we found out that the reactor pressure is highly correlated with the product separator pressure (xmeas 13). Therefore, we constructed prediction model for the separator pressure using $xmv6$ and $xmeas13$ as inputs and we define a multi-variable block. Figure 12 illustrates how an stealthy attack in the reactor pressure is rapidly detected by D_{7-13} due to

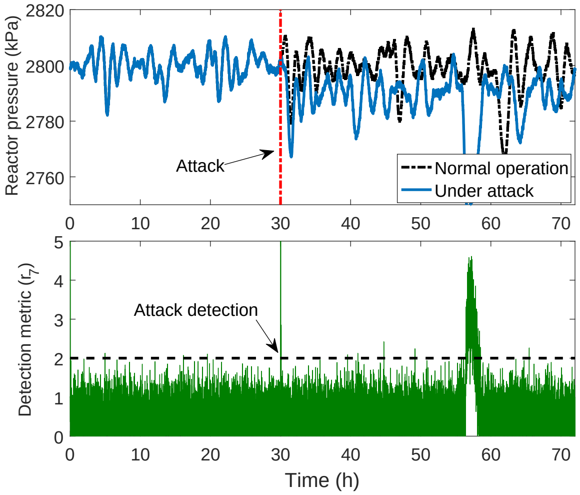


Figure 11: Reactor pressure and detection metric when an bias attack of the form $y_7^a = y_7 + 10$ is launched after 30 h. Notice that the sudden changes make the detection metric to grow rapidly and detect the attack.

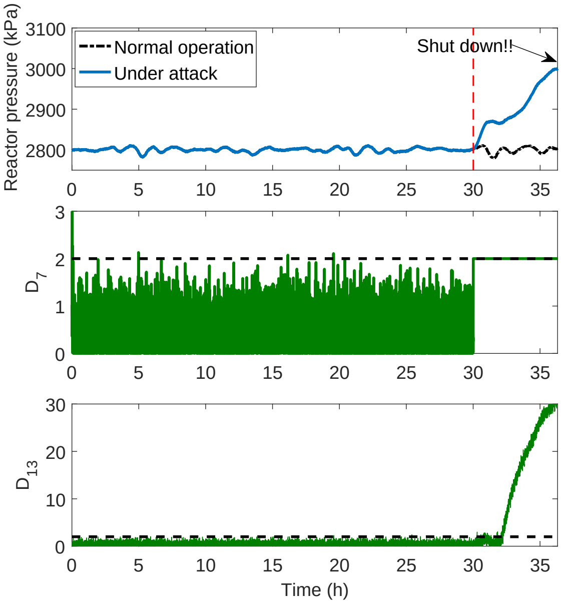


Figure 12: MVD when an optimal stealthy attack is launched only in sensor xmeas 7 (y_7). The attack remains stealthy for the detection D_7 even though the reactor pressure grows until it reaches unsafe states. However, due to the high correlation between y_7 and y_{13} (i.e., $corr(s_7, s_{13}) = 0.96$), D_{7-13} is able to detect the attack.

the high correlation between the signals. Similarly, other blocks can be constructed by using the stripper pressure sensor and other correlated measurements. As a consequence, an adversary will have to compromise all the correlated sensors in order to remain completely stealthy. Even if an attacker gains access to both sensors, the stealthy attack will not have a damaging impact in the system as depicted in Figure 13. Clearly, adding sensors is able to limit the impact of powerful attackers.

As an additional measure of security, we implemented the proposed mitigation strategy such that if an attack is detected, the compromised sensor reading is replaced by an estimation obtained from one of its correlated sensors. In our case study, the reactor pressure is highly correlated with the stripper pressure. Therefore, using an autoregressive model we can estimate the reactor pressure from the stripper pressure. Figure 14 illustrates how our proposed mitigation strategy is able to ensure the stable operation of the system even in the presence of an attack.

5.1 Comparing Detection Architectures

In order to compare how different detection mechanisms perform depending on the amount of redundant sensors included and on the type of architecture, we use the **evaluation metric for the effectiveness of physics-based anomaly detection** introduced in [22]. This metric takes into account the usability and security factors by analyzing the trade-off between the impact of the worst attack the adversary can launch while remaining undetected (y-axis) and the average time between false alarms (x-axis).

Y-axis (Security). The adversary wants to drive the system to the worst possible condition it can without being detected, where ‘'worst” refers to *the maximum deviation of a signal from its true value that the attacker can obtain* (without raising an alarm, and given a fixed-period of time, otherwise given infinite time, the attacker might be able to grow this deviation without bound).

X-axis (Usability). Typically, the false alarm rate is used to measure the usability of a detection mechanism. However, it has been shown that using the expected time between false alarms $E[T_{fa}]$ offers several advantages. The usability metric can be computed by counting the number of false alarms nF_A for an experiment with a duration T_E under normal operation (without attack) for several τ. Then, for each τ we calculate the estimated time for a false alarm by $E[Tfa] = T_E/nF_A$.
As a consequence, small τ will cause small $E[Tfa]$, but it will limit the impact that an adversary can cause in order to remain stealthy.

We launched stealthy attacks for different detection strategies that involve single and multiple sensors. Figure 15 depicts the reactor pressure increase for each stealthy attack. Without attack, the reactor pressure is 2800 kPa. If it reaches 3000 kPa, the plant is shut down. Using only one sensor allows the adversary to drive the reactor pressure to dangerous values; however, using MVD with highly correlated sensors limit significantly what the attacker can

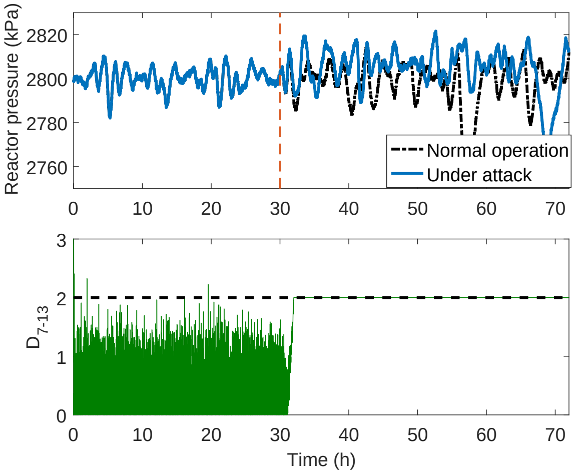


Figure 13: MVD when an optimal stealthy attack is launched in sensor y_7 with the detection D_{7-13}. Even though the attack is stealthy it causes a small deviation from the nominal operation. The attack might only be detected if we consider the correlation with other control loops.

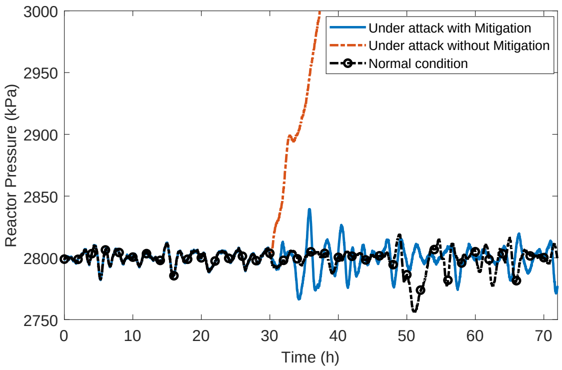


Figure 14: Reactor pressure for 3 different cases. Notice that with the proposed mitigation strategy, the system continues operating close to nominal conditions.

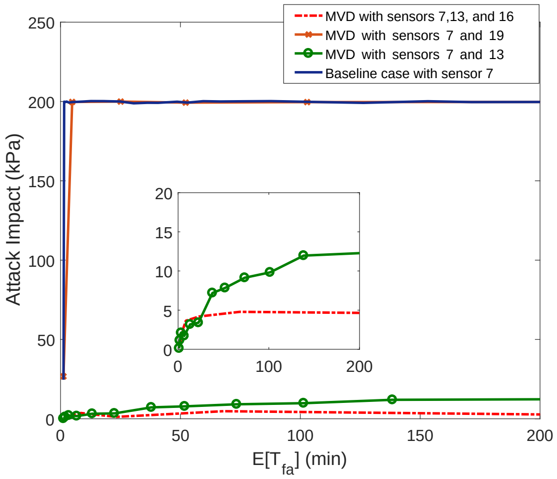


Figure 15: Comparison between different detection strategies. Sensors 13 and 16 are highly correlated with sensor 7, but sensor 19 is not. Notice that including correlated sensors in the prediction models significantly decreases the impact of stealthy attacks. On the other hand, a stealthy attack for the baseline case and by including sensor 19 is able to drive the system to dangerous pressure levels causing the plant to shut down.

do. On the other hand, when a MVD is constructed with a non-correlated sensor (sensor 19), it cannot prevent the shut down of the plant.

6 CONCLUSIONS

In this paper we have introduced the concept of hidden sensors: the idea of deploying new sensors to measure variables that are not being used by the process, in order to use them to improve the security of cyber-physical systems. We showed how these new measurements can be identified (correlation analysis) and then deployed for better attack-detection (MVD) in Figure 3 and better attack mitigation (by replacing the attacked-sensor with the new hidden sensor estimates).

We implemented our attacks and defenses in a realistic Industrial Control testbed controlling the TE process with classical industrial technologies and industrial network protocols representative of protocols that have been attacked in the real-world [5, 23].

Our results show that attacks that are not detected by previous proposals can be detected by our new hidden variables, as shown at the bottom of Figure 12. In addition we show that if the attacker becomes aware of our hidden variables, and tries to launch an attack that bypasses our hidden-variable anomaly detection algorithm, then it will not succeed in driving the plant to an unsafe state, as illustrated in Figure 13. Finally, our attack-mitigation strategy (replacing the compromised sensor value with an estimate given by the hidden sensor) is also able to keep the system safe under attack, as illustrated in Figure 14.

7 ACKNOWLEDGEMENTS

This work was performed under the financial assistance award 70NANB17H282N from U.S. Department of Commerce, National Institute of Standards and Technology (NIST).

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. This publication was co-authored by United States Government employees as part of their official duties and is, therefore, a work of the U.S. Government and not subject to copyright. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NIST.

REFERENCES

[1] Chuadhry Mujeeb Ahmed, Carlos Murguia, and Justin Ruths. 2017. Model-based attack detection scheme for smart water distribution networks. In *Asia Conference on Computer and Communications Security (AsiaCCS)*. ACM, 101–113.

[2] Wissam Aoudi, Mikel Iturbe, and Magnus Almgren. 2018. Truth Will Out: Departure-Based Process-Level Detection of Stealthy Attacks on Control Systems. In *Conference on Computer and Communications Security (CCS)*. ACM, 817–831.

[3] Open DeviceNet Vendors Association et al. 2013. The CIP networks Library, Volume 5, CIP Safety. (2013).

[4] Alvaro A Cárdenas, Saurabh Amin, Zong-Syun Lin, Yu-Lun Huang, Chi-Yen Huang, and Shankar Sastry. 2011. Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the 6th ACM symposium on information, computer and communications security*. ACM, 355–366.

[5] Anton Cherepanov. 2017. WIN32/INDUSTROYER: A new threat for industrial control systems. *White paper, ESET (June 2017)* (2017).

[6] Hongjun Choi, Wen-Chuan Lee, Yousra Aafer, Fan Fei, Zhan Tu, Xiangyu Zhang, Dongyan Xu, and Xinyan Xinyan. 2018. Detecting Attacks Against Robotic Vehicles: A Control Invariant Approach. In *Conference on Computer and Communications Security (CCS)*. ACM, 801–816.

[7] James J Downs and Ernest F Vogel. 1993. A plant-wide industrial process control problem. *Computers & chemical engineering* 17, 3 (1993), 245–255.

[8] Helen Durand. 2018. A nonlinear systems framework for cyberattack prevention for chemical process control systems. *Mathematics* 6, 9 (2018), 169.

[9] Sriharsha Etigowni, Dave Jing Tian, Grant Hernandez, Saman Zonouz, and Kevin Butler. 2016. CPAC: securing critical infrastructure with cyber-physical access control. In *Annual Computer Security Applications Conference (ACSAC)*. ACM, 139–152.

[10] Kevin Fu and Wenyuan Xu. 2018. Risks of trusting the physics of sensors. *Commun. ACM* 61, 2 (2018), 20–23.

[11] Jairo Giraldo, David Urbina, Alvaro Cardenas, Junia Valente, Mustafa Faisal, Justin Ruths, Nils Ole Tippenhauer, Henrik Sandberg, and Richard Candell. 2018. A Survey of Physics-Based Attack Detection in Cyber-Physical Systems. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 76.

[12] IETF Network Working Group. 2015. Administratively Scoped IP Multicast. (October 2015). http://tools.ietf.org/html/rfc2365.

[13] Dina Hadžiosmanović, Robin Sommer, Emmanuele Zambon, and Pieter H Hartel. 2014. Through the eye of the PLC: semantic security monitoring for industrial processes. In *Annual Computer Security Applications Conference (ACSAC)*. ACM, 126–135.

[14] Yu-Lun Huang, Alvaro A Cárdenas, Saurabh Amin, Zong-Syun Lin, Hsin-Yi Tsai, and Shankar Sastry. 2009. Understanding the physical and economic consequences of attacks on control systems. *International Journal of Critical Infrastructure Protection* 2, 3 (2009), 73–83.

[15] Ali Khaki-Sedigh and Bijan Moaveni. 2009. *Control Configuration Selection of Nonlinear Multivariable Plants*. Springer Berlin Heidelberg, Berlin, Heidelberg, 139–172.

[16] Denis Foo Kune, John Backes, Shane S Clark, Daniel Kramer, Matthew Reynolds, Kevin Fu, Yongdae Kim, and Wenyuan Xu. 2013. Ghost talk: Mitigating EMI signal injection attacks against analog sensors. In *Symposium on Security and Privacy (S&P)*. IEEE, 145–159.

[17] Yao Liu, Peng Ning, and Michael K Reiter. 2009. False data injection attacks against state estimation in electric power grids. In *Conference on Computer and Communications Security (CCS)*. ACM, 21–32.

[18] Bijan Moaveni and Ali Khaki-Sedigh. 2007. Input-output pairing for nonlinear multivariable systems. *Journal of applied sciences* 7, 22 (2007), 3492–3498.

[19] N Lawrence Ricker. 1996. Decentralized control of the Tennessee Eastman challenge process. *Journal of Process Control* 6, 4 (1996), 205–221.

[20] Timothy Trippel, Ofir Weisse, Wenyuan Xu, Peter Honeyman, and Kevin Fu. 2017. WALNUT: Waging doubt on the integrity of MEMS accelerometers with acoustic injection attacks. In *European Symposium on Security and Privacy (EuroS&P)*. IEEE, 3–18.

[21] David Urbina, Yufei Gu, Juan Caballero, and Zhiqiang Lin. 2014. SigPath: A memory graph based approach for program data introspection and modification. In *European Symposium on Research in Computer Security*. Springer, 237–256.

[22] David I Urbina, Jairo A Giraldo, Alvaro A Cardenas, Nils Ole Tippenhauer, Junia Valente, Mustafa Faisal, Justin Ruths, Richard Candell, and Henrik Sandberg. 2016. Limiting the impact of stealthy attacks on industrial control systems. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1092–1105.

[23] Kyle Wilhoit. 2014. Havex, it is down with OPC. *Threat Research| FireEye Inc. Retrieved July* 11 (2014), 2015.

SEISMIC STABILITY ASSESSMENT OF STEEL MOMENT FRAMES AND IMPLICATIONS FOR DESIGN

L.A. Fahnestock[(1)], S. Shi[(2)], M.S. Speicher[(3)]

[(1)] *Professor, Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, fhnstck@illinois.edu*
[(2)] *Graduate Research Assistant, Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, shi40@illinois.edu*
[(3)] *Research Structural Engineer, Earthquake Engineering Group, National Institute of Standards and Technology, speicher@nist.gov*

Abstract

Although it is clear that a building must be capable of carrying gravity loads while developing large inelastic deformations and associated lateral displacements during a large earthquake, achieving this performance objective in day-to-day practice still represents a major challenge. Current code-based consideration of seismic P-Delta effects is generally based on simplistic elastic models, and despite major advances in seismic systems and analysis techniques, no simple and reliable design methods for seismic stability are available.

Specific to steel buildings and the design framework in the United States, the current fundamental approach for stability design was developed and calibrated for non-seismic scenarios where the structure has modest overstrength and the ultimate strength (stability point) of the structure occurs prior to significant inelastic deformation. However, in a ductile steel seismic lateral force-resisting system (LFRS), the design-level forces and resulting nominally-elastic deformations are not consistent with the ultimate strength state of the system, which corresponds to significant overstrength and inelastic deformation.

Despite the vastly different behaviors expected in wind-dominated design vs. seismic-dominated design, the same stability design approach is employed. This stability design approach is nominally based on second-order elastic analysis (i.e., in the structural analysis model, equilibrium is formulated on the elastic deformed position and inelastic response is not considered). However, in seismic design it is not rational to consider P-Delta effects at elastic deformation levels.

The results described in this paper are part of a comprehensive study that is seeking to identify the most critical LFRS parameters that affect seismic stability and to develop a rigorous yet simple methodology whereby these parameters can be considered in design. This paper focuses on a set of steel special moment frames that is designed with or without consideration of stiffness reduction due to inelasticity, elastic P-Delta effects and drift limits. The moment frame designs are interrogated using nonlinear static and dynamic analyses to assess their collapse potential and to identify the most important parameters for design. The results from this paper will be combined with similar assessments for other types of steel seismic LFRS to propose design provisions that will enhance safety and economy for future design.

Keywords: Steel Structures; Seismic Stability; Design Provisions

Fahnestock, Larry; Shi, Shitao; Speicher, Matthew. "Seismic Stability Assessment of Steel Moment Frames and Implications for Design." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Seismic stability design has been studied for many years [1-3], and although modern performance-based seismic design does employ advanced analysis to rigorously consider stability effects, there has been relatively little advancement in how seismic stability is considered in pragmatic code-based procedures. In the United States, most buildings are designed for seismic loads using reduced forces based on the R factor prescribed in *Minimum Design Loads and Associated Criteria for Buildings and Other Structures* (ASCE/SEI 7-16) [4] with second-order elastic analysis – namely, the strength of the lateral force-resisting system (LFRS) is determined considering elastic P-Delta effects. However, research indicates that parameters related to inelastic response – such as post-yield stiffness and overstrength – are more important for seismic stability than initial elastic strength (i.e., the system yield strength).

Focusing on steel buildings and the *Specification for Structural Steel Buildings* (ANSI/AISC 360-16) [5], stability design is conducted with the direct analysis method (DM), which was developed and calibrated for non-seismic scenarios where the structure has modest overstrength and the ultimate strength (stability point) of the structure occurs prior to significant inelastic deformation. In a ductile steel seismic LFRS, such as a special moment frame (SMF), the design-level forces and resulting nominally-elastic deformations are not consistent with the ultimate strength state of the system, which corresponds to significant overstrength and inelastic deformation. Although the *Seismic Provisions for Structural Steel Buildings* (ANSI/AISC 341-16) [6] contain rigorous requirements related to capacity-based proportioning and ductile detailing, global seismic stability is not directly considered. The code-based seismic drift limit per ASCE 7 [4] is the primary means by which seismic stability is indirectly considered in the design process. In describing the seismic design landscape two decades ago, Gupta and Krawinkler [3] wrote, “At this time, no simple procedure can be recommended that will permit a definite assessment of collapse hazard due to P-Delta effects.” Current code-based consideration of seismic stability has advanced little since that time, and a more fundamental approach is needed.

Several studies have provided preliminary insight into the effectiveness of current seismic design and assessment approaches. Buckling-restrained braced frame (BRBF) and BRBF-SMF dual systems were examined, and the study indicated that ignoring stiffness reduction, imperfections and (elastic) P-Delta effects in design while considering post-yield stiffness can provide acceptable seismic performance [7]. Furthermore, an earlier study examined SMFs and the effects of residual stresses and imperfections were shown to be unimportant for seismic response [8]. In a study that was focused on consistency between ASCE 7 and ASCE 41 [9] for BRBF, SMF, special concentrically-braced frame (SCBF) and eccentrically-braced frame (EBF) systems, nonlinear models were used to assess the performance of code-compliant buildings [10-14]. In general, the results indicate that the ASCE 41 assessment procedure is overly conservative and that collapse performance evaluated with the *Quantification of Building Seismic Performance Factors* FEMA P695 methodology [15] is generally acceptable for buildings designed per ASCE 7.

This paper summarizes the initial portion of a study that is comprehensively evaluating seismic stability design of steel frames. A series of code-compliant and non-code-compliant SMFs were designed and evaluated with nonlinear static (pushover) and nonlinear dynamic (response history) analyses. Differences in observed response are discussed and implications for future work are presented.

2. Building Designs and Numerical Models

2.1 Prototype Building Designs

The 8-story and 16-story office buildings referenced in NIST Technical Note 1863-1 [16] were the basis of the prototype buildings designed for this research. As shown in Fig. 1 for the 8-story building, the rectangular plan of the prototype buildings contains two perimeter 3-bay SMFs in the East-West direction and four perimeter SCBFs in the North-South direction. The research reported here focuses on the SMFs only, and the designs are based on ASCE 7-16, which leads to small variations compared to the original

NIST designs that were based on ASCE 7-05. All other parameters for the prototype building are consistent with NIST Technical Note 1863-1 [16]. Along with ASCE 7-16, AISC 341-16 and AISC 360-16 were used to design the prototype buildings. Stability was considered in accordance with the Direct Analysis Method (DM) in Ch. C of AISC 360-16 [5], which includes elastic P-Delta effects and stiffness reduction. There are two basic approaches for seismic lateral force analysis in ASCE 7-16: the equivalent lateral force procedure (ELF) and modal response spectrum analysis (RSA). ELF approximates a first-mode force profile, whereas RSA includes contributions from multiple modes. In both approaches, the analyses are elastic (no material nonlinearity) but do included geometric nonlinearity (P-Delta effects).

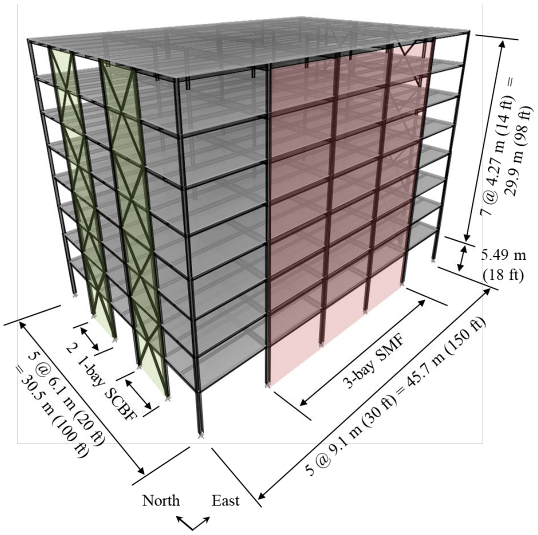
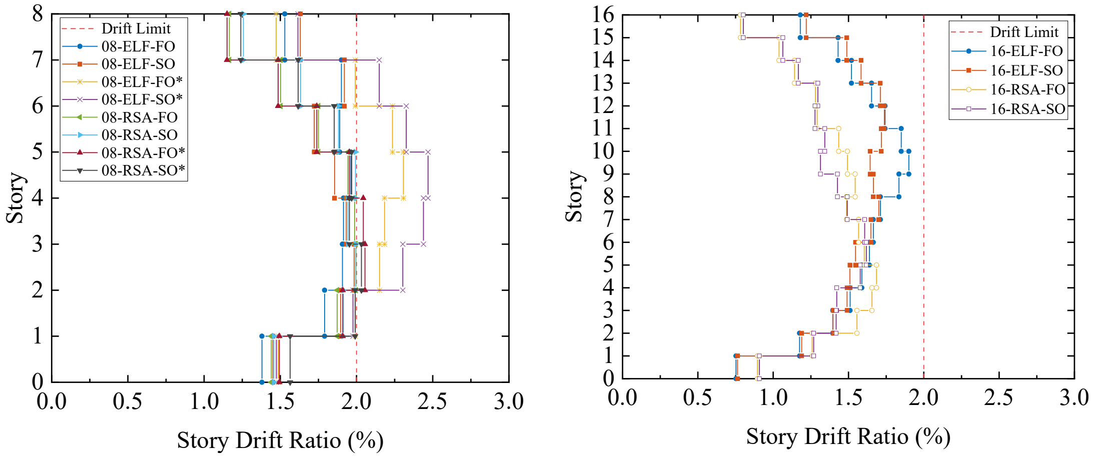

7 @ 4.27 m (14 ft) = 29.9 m (98 ft)
5.49 m (18 ft)
3-bay SMF
5 @ 9.1 m (30 ft) = 45.7 m (150 ft)
2 1-bay SCBF
5 @ 6.1 m (20 ft) = 30.5 m (100 ft)
North
East

Fig. 1 – 8-story Prototype Building

The frame design matrix is based on code-compliant designs that use second-order elastic analysis. From this baseline design, variations are made to study the effect of analysis type (first-order elastic) and the effect of a seismic drift limit. Table 1 shows the design matrix, where FO represents a non-code-compliant design ignoring stiffness reduction and P-Delta effects, and SO represents a code-compliant design. An asterisk indicates a design that ignores the drift limit. Table 1 indicates that the controlling requirement of all 16-story SMFs was strength (and drift was satisfied automatically). A summary of frame member sections is provided in Tables 2 and 3, and seismic design parameters are presented in Table 4. The dead load calculation considered self-weight for each design, so seismic weights (D + 0.2L) are slightly different within each building scenario. V_{design} is the base shear for strength design and V_{drift} is the base shear for the drift check. For reference, the story drift profiles under drift check lateral forces are shown in Fig. 2.

Table 1 – Design Matrix of Special Moment Frames

Analysis Type	Drift Check	8-Story		16-Story	
		ELF	RSA	ELF	RSA
First-Order	Yes	08-ELF-FO	08-RSA-FO	16-ELF-FO	16-RSA-FO
Second-Order	Yes	08-ELF-SO	08-RSA-SO	16-ELF-SO	16-RSA-SO
First-Order	No	08-ELF-FO*	08-RSA-FO*	–	–
Second-Order	No	08-ELF-SO*	08-RSA-SO*	–	–



(a)

(b)

Fig. 2 – Design Story Drift Ratio Profiles: (a) 8-story SMFs; (b) 16-story SMFs

Table 2 – 8-Story Special Moment Frame Designs

Level	Beam	Interior Column	Exterior Column	Beam	Interior Column	Exterior Column
		8-ELF-FO			8-ELF-SO	
8th /Roof	W 24 × 55	W 18 × 143	W 18 × 86	W 24 × 55	W 18 × 143	W 18 × 86
6th /7th	W 27 × 94	W 18 × 234	W 18 × 119	W 30 × 108	W 18 × 258	W 18 × 143
4th /5th	W 30 × 108	W 18 × 258	W 18 × 143	W 30 × 116	W 18 × 283	W 18 × 175
2nd / 3rd	W 30 × 108	W 18 × 258	W 18 × 192	W 30 × 116	W 18 × 283	W 18 × 211
		8-ELF-FO*			8-ELF-SO*	
8th /Roof	W 24 × 55	W 18 × 143	W 18 × 86	W 24 × 55	W 18 × 143	W 18 × 86
6th /7th	W 24 × 76	W 18 × 192	W 18 × 106	W 24 × 84	W 18 × 211	W 18 × 143
4th /5th	W 27 × 94	W 18 × 234	W 18 × 143	W 27 × 94	W 18 × 234	W 18 × 158
2nd /3rd	W 27 × 94	W 18 × 234	W 18 × 192	W 30 × 108	W 18 × 258	W 18 × 192
		8-RSA-FO			8-RSA-SO	
8th /Roof	W 24 × 55	W 18 × 143	W 18 × 86	W 24 × 55	W 18 × 143	W 18 × 86
6th /7th	W 24 × 76	W 18 × 192	W 18 × 106	W 24 × 84	W 18 × 211	W 18 × 143
4th /5th	W 27 × 94	W 18 × 234	W 18 × 143	W 27 × 94	W 18 × 234	W 18 × 158
2nd / 3rd	W 27 × 94	W 18 × 234	W 18 × 192	W 30 × 108	W 18 × 258	W 18 × 192
		8-RSA-FO*			8-RSA-SO*	
8th /Roof	W 21 × 44	W 18 × 119	W 18 × 55	W 21 × 44	W 18 × 119	W 18 × 55
6th /7th	W 24 × 55	W 18 × 143	W 18 × 65	W 24 × 55	W 18 × 143	W 18 × 86
4th /5th	W 24 × 62	W 18 × 175	W 18 × 97	W 24 × 76	W 18 × 192	W 18 × 106
2nd / 3rd	W 24 × 76	W 18 × 192	W 18 × 130	W 24 × 84	W 18 × 211	W 18 × 158

Table 3 – 16-Story Special Moment Frame Designs

Level	Beam	Interior Column	Exterior Column	Beam	Interior Column	Exterior Column
	16-ELF-FO (16-ELF-FO*)			16-ELF-SO (16-ELF-SO*)		
16th /Roof	W 24 × 55	W 27 × 129	W 27 × 94	W 24 × 55	W 27 × 129	W 27 × 94
14th /15th	W 27 × 94	W 27 × 235	W 27 × 114	W 27 × 94	W 27 × 235	W 27 × 114
12th /13th	W 30× 108	W 27 × 258	W 27 × 129	W 30× 108	W 27 × 258	W 27 × 194
10th /11th	W 30× 108	W 27 × 258	W 27 × 194	W 33× 130	W 27 × 307	W 27 × 235
8th /9th	W 33× 130	W 27 × 307	W 27 × 194	W 33× 130	W 27 × 307	W 27 × 235
6th /7th	W 33× 130	W 27 × 307	W 27 × 258	W 33× 152	W 27 × 368	W 27 × 336
4th /5th	W 33× 130	W 27 × 368	W 27 × 307	W 33× 152	W 27 × 368	W 27 × 336
2nd / 3rd	W 33× 130	W 27 × 368	W 27 × 539	W 33× 152	W 27 × 368	W 27 × 539
	16-RSA-FO (16-RSA-FO*)			16-RSA-SO (16-RSA-SO*)		
16th /Roof	W 24 × 55	W 27 × 129	W 27 × 94	W 24 × 55	W 27 × 129	W 27 × 94
14th /15th	W 24 × 76	W 27 × 178	W 27 × 94	W 24 × 76	W 27 × 235	W 27 × 102
12th /13th	W 27 × 94	W 27 × 235	W 27 × 102	W 27 × 94	W 27 × 235	W 27 × 102
10th /11th	W 27 × 94	W 27 × 235	W 27 × 102	W 30× 108	W 27 × 258	W 27 × 129
8th /9th	W 30× 108	W 27 × 258	W 27 × 129	W 30× 108	W 27 × 258	W 27 × 129
6th /7th	W 30× 108	W 27 × 258	W 27 × 129	W 30× 108	W 27 × 258	W 27 × 146
4th /5th	W 30× 108	W 27 × 258	W 27 × 146	W 33× 130	W 27 × 307	W 27 × 161
2nd / 3rd	W 33× 130	W 27 × 307	W 27 × 217	W 33× 130	W 27 × 307	W 27 × 235

Table 4 – Summary of Seismic Design Parameters for Prototype Building in East-West direction

Case	W_{total} (kips)	W_{SMF} (kips)	V_{design} (kips)	V_{drift} (kips)	C_uT_a (s)	T_1 (s)
08-ELF-FO	10651	299	513	333	1.76	2.60
08-ELF-SO	10679	327	514	338	1.76	2.42
08-ELF-FO*	10623	271	512	–	1.76	2.97
08-ELF-SO*	10642	290	513	–	1.76	2.77
08-RSA-FO	10566	211	509	240	1.76	3.80
08-RSA-SO	10581	219	509	230	1.76	3.53
08-RSA-FO*	10555	201	509	–	1.76	3.82
08-RSA-SO*	10573	218	509	–	1.76	3.58
16-ELF-FO	21828	788	958	442	3.02	3.90
16-ELF-SO	21897	860	961	447	3.02	3.63
16-RSA-FO	21652	608	951	372	3.02	4.82
16-RSA-SO	21679	638	953	358	3.02	4.54

2.2 Numerical Models

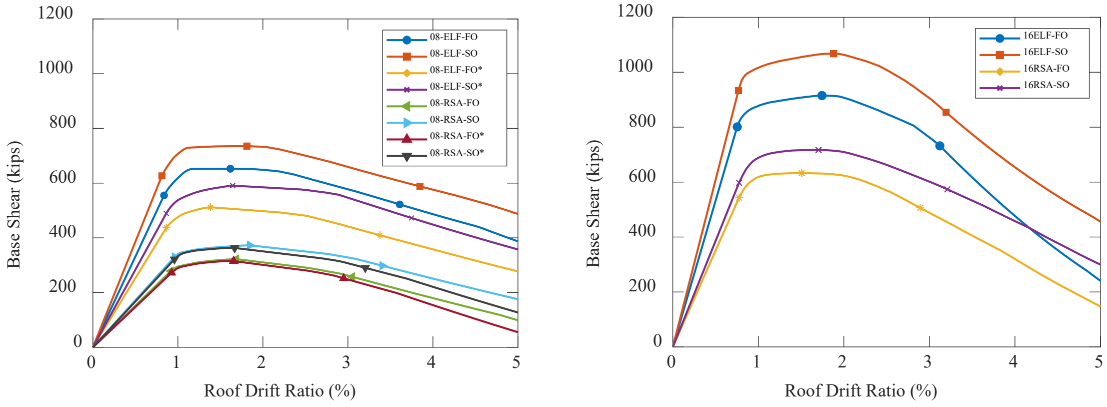
The OpenSees framework [17] was used to carry out the nonlinear static (pushover) analyses and nonlinear dynamic (response history) analyses of the SMF models. Given that the lateral force-resisting frames are located at the perimeter and the building is symmetric, the numerical models were 2D. Modal damping of 3% was used and an additional 0.3% stiffness-proportional damping was applied to damp out higher modes. Nonlinear rotational springs were placed at the ends of the SMF columns (half the column depth from the face of the beam) and at the center of each reduced beam section (RBS) connection. The nonlinear springs use the modified Ibarra Medina Krawinkler (IMK) deterioration model, which simulates in-cycle and cyclic degradation [18]. The force-deformation parameters for the RBS connections followed the recommendations made by Lignos and Krawinkler [19], which were derived using multivariate regression analysis of a database of experimental results. The force-deformation parameters for the column hinges followed the recommendations produced by NIST [20] using a monotonic backbone. The panel zones were modeled using the approach outlined by Krawinkler [21]. Additionally, although the column splice was designed at 1.2 m (4 ft) above the beam-to-column joint, this was ignored in the model and section size changes were made at the floor levels.

To approximately capture the behavior of the gravity framing system, a leaning column was used. The leaning column was assigned a moment of inertia equal to the sum of the moments of inertia of the tributary gravity frame columns and SCBF columns. The leaning column was attached to each floor of the SMF using equal degree of freedom constraints in the horizontal direction. Additionally, elastic-plastic hinges were placed at the top and bottom of each story and assigned a strength equal to the sum of the plastic moments of the non-SMF columns.

3. Numerical Simulations

3.1 Nonlinear Static Analyses

Nonlinear static analyses were performed to evaluate and compare behavior of the prototype designs. Based on the nonlinear analysis procedures described in FEMA P695 [15], the gravity loading applied in the numerical models was 1.05D + 0.25L, and the lateral load distribution was in proportion to the fundamental mode shape. Pushover curves (base shear vs. roof drift ratio) are shown in Fig. 3, with the following three important points marked: a) end of the linear range, b) peak base shear (V_{max}), and c) $0.8V_{max}$. These curves demonstrate that all frames exhibit significant softening around 1% roof drift ratio and reach V_{max} between 1% and 2% roof drift ratio. Beyond V_{max}, significant negative stiffness develops due to the global P-Delta effects. Response quantities from the pushover analyses provide useful comparisons between the prototype designs. Ultimate displacement (δ_u) is defined as the roof displacement at $0.8V_{max}$, and the effective yield displacement ($\delta_{y,eff}$) is calculated per FEMA P695 [15] and used as a reference for defining ductility as $\mu = \delta_u / \delta_{y,eff}$. System overstrength is defined as $\Omega = V_{max} / V_{design}$. A summary of these response quantities is provided in Table 5.



(a) 8-story SMFs

(b) 16-story SMFs

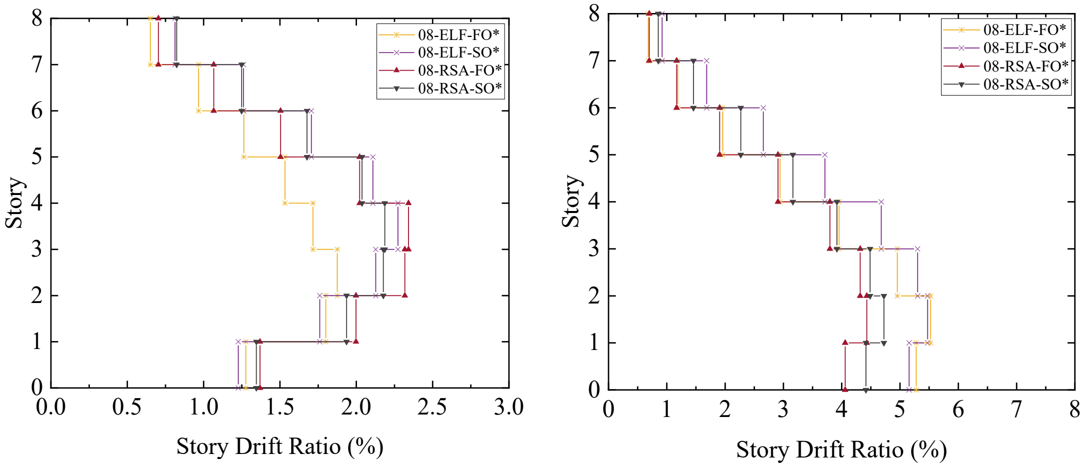
Fig. 3 – Pushover Curves

Fahnestock, Larry; Shi, Shitao; Speicher, Matthew. "Seismic Stability Assessment of Steel Moment Frames and Implications for Design." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

Table 5 – Summary of Response Quantities of One Perimeter Frame from Pushover Analyses

Case	V_{design} (kips)	V_{drift} (kips)	V_{max} (kips)	$\delta_{y,eff}$ / h (%)	δ_u / h (%)	Ω	μ	K_1 (k/in)
08-ELF-FO	256	167	653	0.98	3.61	2.54	3.67	47.6
08-ELF-SO	257	169	735	0.95	3.85	2.86	4.04	55.4
08-ELF-FO*	256	–	512	1.01	3.38	2.00	3.34	36.3
08-ELF-SO*	256	–	591	1.04	3.75	2.31	3.60	40.7
08-RSA-FO	254	120	322	1.06	3.05	1.27	2.88	21.8
08-RSA-SO	255	115	373	1.08	3.40	1.46	3.14	24.7
08-RSA-FO*	254	–	315	1.07	2.95	1.24	2.76	21.1
08-RSA-SO*	254	–	363	1.08	3.20	1.43	2.96	24.1
16-ELF-FO	479	221	915	0.86	3.12	1.91	3.63	38.9
16-ELF-SO	481	224	1068	0.88	3.20	2.22	3.64	44.4
16-RSA-FO	476	186	633	0.90	2.89	1.33	3.21	25.6
16-RSA-SO	476	179	717	0.93	3.21	1.50	3.46	28.2

As shown in Table 5, V_{design} is essentially equal for designs of the same building using ELF and RSA, but Ω is significantly reduced for RSA designs compared to ELF designs. For the two code-compliant ELF designs, Ω is greater than 2, whereas for the two code-compliant RSA designs, Ω is around 1.5. For the code-compliant designs, the 8-story ELF design has approximately 30% greater μ than the RSA design, whereas the 16-story ELF design has only 5% greater μ than the RSA design. The impact of these differences in overstrength and ductility, in conjunction with other parameters, requires further investigation through dynamic analysis.

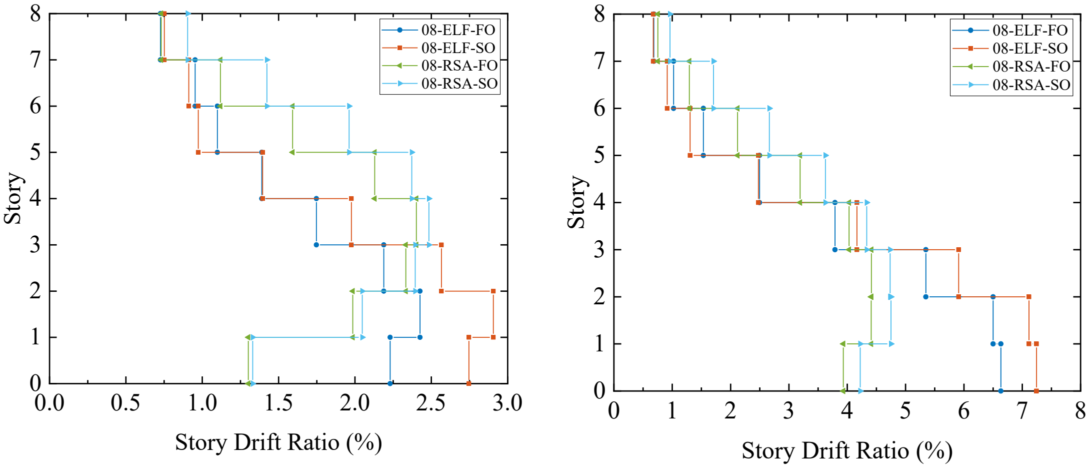


(a) At Maximum Base Shear Capacity (V_{max})

(b) At Ultimate Displacement (δ_u)

Fig. 4 – Effect of Analysis Type on Story Drift Profile of 8-story SMFs without Drift Limit

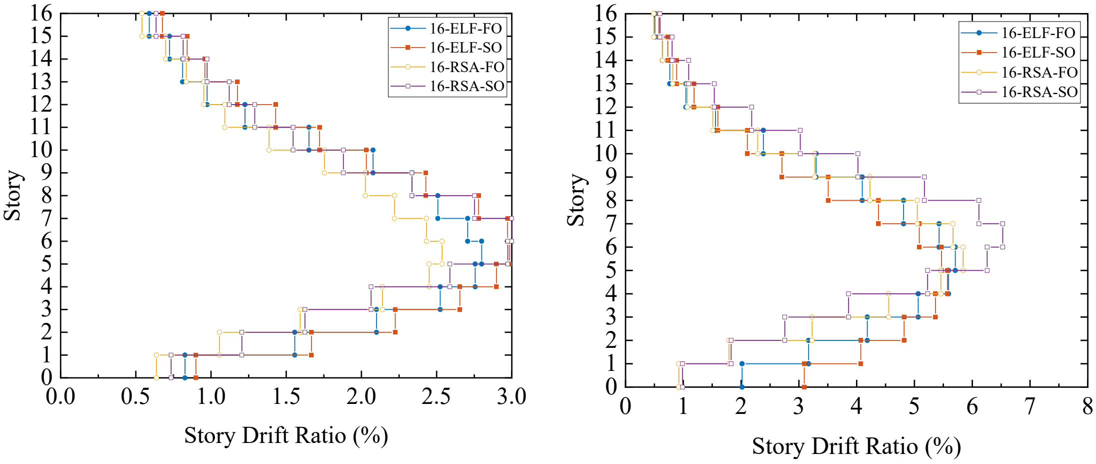
The influences of analysis type on the distributions of inelasticity in the pushover analyses are presented in the story drift profiles taken at the points of V_{max} and δ_u, as shown in Fig. 4, Fig. 5 and Fig. 6. These figures demonstrate that story drift was generally concentrated in the middle stories at the points of maximum base shear, and then drift became more pronounced in the lower stories in the region of global negative stiffness. The shape of the story drift profiles of SMFs designed considering elastic P-Delta effects are similar to the cases designed without elastic P-Delta effects. Comparison of the 8-story ELF designs with and without drift limits indicates that stiffening of the frame in the middle stories to meet drift limits leads to greater concentration of inelastic demands in the lower stories. However, this is not observed in the 8-story RSA designs, which consider higher modes of response.



(a) At Maximum Base Shear Capacity (V_{max})

(b) At Ultimate Displacement (δ_u)

Fig. 5 – Effect of Analysis Type on Story Drift Profile of 8-story SMFs from Pushover Analyses



(a) At Maximum Base Shear Capacity (V_{max})

(b) At Ultimate Displacement (δ_u)

Fig. 6 – Effect of Analysis Type on Story Drift Profile of 16-story SMFs from Pushover Analyses

The effects of drift limit and analysis type on response quantities are summarized in Table 6 and Table 7, respectively, by calculating the ratios of response quantities. Comparing cases with and without drift limit for the 8-story designs, removing the drift limit is seen to reduce the weight of steel in the special moment frame (W_{SMF}) by 11% in the code-compliant ELF design, but only 1% in the code-compliant RSA design. Correspondingly, V_{max} was reduced by 20% in the ELF design and only 3% in the RSA design. The drift limit has a significant impact on initial stiffness for the ELF design (over 30%), but minimal impact for the RSA design (less than 5%).

Considering the impact of P-Delta effects (second-order analysis), which also includes stiffness reduction per the Direct Analysis Method (DM), Table 7 shows an increase of W_{SMF} roughly in the range of 5-10%. This increased steel weight translates into increases in strength of roughly 10-15% and increases in ductility of roughly 5-10%. The pushover curves in Fig. 3 illustrate these increases graphically, and the increases in elastic stiffness are also evident. Elastic stiffness increases are approximately on the same order as the strength increases (10-15%).

Table 6 – Effect of Drift Limit on Response Quantities from Pushover Analyses

Case	$\frac{(W_{SMF})^*}{(W_{SMF})}$	$\frac{(V_{max})^*}{(V_{max})}$	$\frac{(\Omega)^*}{(\Omega)}$	$\frac{(\mu)^*}{(\mu)}$	$\frac{(K_1)^*}{(K_1)}$
08-ELF-FO*/08-ELF-FO	0.91	0.78	0.79	0.91	0.76
08-ELF-SO*/08-ELF-SO	0.89	0.80	0.81	0.89	0.73
08-RSA-FO*/08-RSA-FO	0.95	0.98	0.98	0.96	0.97
08-RSA-SO*/08-RSA-SO	0.99	0.97	0.97	0.94	0.97

Table 7 – Effect of Analysis Type on Response Quantities from Pushover Analyses

Case	$\frac{(W_{SMF})_{FO}}{(W_{SMF})_{SO}}$	$\frac{(V_{max})_{FO}}{(V_{max})_{SO}}$	$\frac{(\Omega)_{FO}}{(\Omega)_{SO}}$	$\frac{(\mu)_{FO}}{(\mu)_{SO}}$	$\frac{(K_1)_{FO}}{(K_1)_{SO}}$
08-ELF-FO/08-ELF-SO	0.91	0.89	0.89	0.91	0.86
08-ELF-FO*/08-ELF-SO*	0.93	0.87	0.87	0.93	0.89
08-RSA-FO/08-RSA-SO	0.96	0.86	0.86	0.92	0.88
08-RSA-FO*/08-RSA-SO*	0.92	0.87	0.87	0.93	0.88
16-ELF-FO/16-ELF-SO	0.92	0.86	0.86	1.00	0.88
16-RSA-FO/16-RSA-SO	0.95	0.88	0.88	0.93	0.91

3.2 Nonlinear Response History Analyses

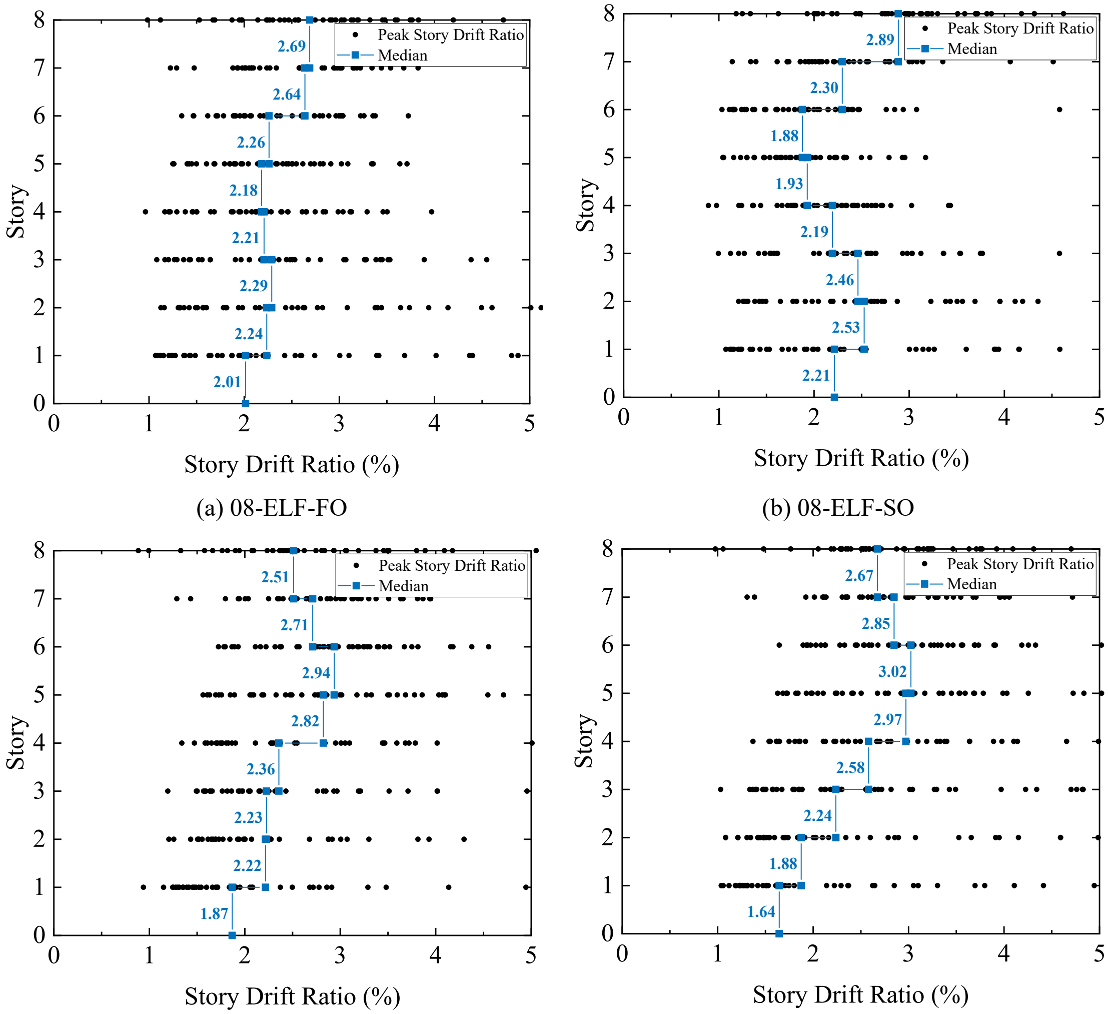
To further evaluate the influence of LFRS design parameters on seismic performance, the 8-story ELF designs were subjected to the far-field record set (44 individual horizontal ground motions) from FEMA P695. The ground motions were normalized per FEMA P695 then the median spectral ordinate of these 44 ground motions scaled to match the acceleration of the risk-targeted MCE response spectrum at the target fundamental period for the building (C_uT_a, tabulated in Table 4).

Table 8 – Median Response Quantities under MCE Ground Motions

Case	**Peak Story Shear (kips)**	**Peak Roof Drift Ratio (%)**	**Peak Story Drift Ratio (%)**
08-ELF-FO	873	1.87	3.14
08-ELF-SO	945	1.77	3.21
08-ELF-FO*	763	2.00	3.35
08-ELF-SO*	859	1.93	3.42

Median response quantities for the set of nonlinear dynamic analyses are summarized in Table 8. For the code-compliant design (08-ELF-SO), the median peak roof drift ratio is 1.77% and the median peak story drift ratio is 3.21%. Considering that the design basis earthquake (DBE) is approximately 2/3 of the MCE, the MCE response indicates a median DBE response around the design target of 2%. Comparing 08-ELF-FO to 08-ELF-SO, the design including P-Delta is seen to have a slightly smaller median peak roof drift ratio, but a slightly larger median peak story drift ratio. This somewhat unusual result arises due to the larger section sizes that are used in the SO design, which lead to redistribution of inelastic response.

The designs without drift limit clearly experience greater inelastic drifts than their counterpart designs with drift limits. In the scenario considered here, the drift limit appears to be more influential than consideration of P-Delta effects. The design by first-order analysis that considers the drift limit (08-ELF-FO) performs better than the design by second-order analysis that ignores the drift limit (08-ELF-SO*). Fig. 7 shows that 08-ELF-FO has the most uniform distribution of drift of the four cases considered. Referring back to Fig. 4 and Fig. 5, the story drift profiles from pushover analyses, which use an elastic first mode force profile, have greater variation. The drift limit, which led to larger section sizes in the middle stories, reduced story drift ratios in this region, but also increased story drift ratios above and below.



(a) 08-ELF-FO

(b) 08-ELF-SO

(c) 08-ELF-FO*

(d) 08-ELF-SO*

Fig. 7 – Median Peak Story Drift Profile under MCE Ground Motion

4. Summary

The ongoing research project described in this paper aims to identify the most critical LFRS parameters that affect seismic stability and to develop a rigorous yet simple methodology whereby these parameters can be considered in design. This paper focuses on a set of steel special moment frames (SMF) that is designed with or without consideration of stiffness reduction due to inelasticity, elastic P-Delta effects and drift limits. Office buildings with two heights (8-story and 16-story) were based on prior work conducted at NIST and designed using current code provisions in the United States. The buildings have regular plan configurations with perimeter lateral force-resisting frames. Equivalent lateral force (ELF) and modal response spectrum analysis (RSA) procedures were used in the design process. The moment frame designs are interrogated using nonlinear static and dynamic analyses to assess their collapse potential and to identify the most important parameters for design. Several observations from the present study are:

- For the 8-story case, removing the drift limit reduces the weight of steel in the SMF by 11% in the code-compliant ELF design, but only 1% in the code-compliant RSA design. From pushover analyses, the peak base shear was reduced by 20% in the ELF design and 3% in the RSA design.
- For the 8-story cases, the drift limit has a significant impact on initial stiffness for the ELF design (over 30%), but minimal impact for the RSA design (less than 5%). For the 16-story cases, the drift limit was not influential since it was satisfied based only on strength requirements.
- Enforcing the drift limit for the 8-story cases led to larger member sizes in the middle stories. This increase in stiffness and strength reduced inelastic dynamic response in stories 4-7, but increased the response above and below.
- Considering the impact of P-Delta effects (second-order analysis), which also includes stiffness reduction per the Direct Analysis Method (DM), the weight of steel in the SMF increased by approximately 5-10% compared to the designs by first-order analysis. This increased steel weight translates into increases in strength of approximately 10-15% and increases in ductility of approximately 5-10%. Elastic stiffness increases are approximately on the same order as the strength increases (10-15%).
- Based on the limited results from this study, the drift limit appears to be more influential than consideration of P-Delta effects.

Further study of seismic stability for steel special moment frames will include comprehensive collapse assessments per FEMA P695. The results from this paper will also be combined with similar assessments for other types of steel seismic lateral force-resisting systems to develop design provisions that will aim to enhance safety and economy for future design.

5. Acknowledgments

This study is supported by the American Institute of Steel Construction. The opinions, findings, and conclusions in this paper are those of the authors and do not necessarily reflect the views of those acknowledged here.

6. Disclaimer

Commercial software may have been used in the preparation of information contributing to this paper. Identification in this paper is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that such software is necessarily the best available for the purpose. NIST policy is to use the International System of Units in all its published materials. However, in this paper, some information is presented in U.S. customary units as this is the preferred system of units in the U.S. earthquake engineering industry.

7. References

[1] MacRae GA (1994): P-Δ effects on single-degree-of-freedom structures in earthquakes. *Earthquake spectra*, **10** (3), 539-568.

[2] Krawinkler H, Seneviratna GD (1998): Pros and cons of a pushover analysis of seismic performance evaluation. *Engineering structures*, **20** (4-6), 452-464.

[3] Gupta A, Krawinkler H (2000): Estimation of seismic drift demands for frame structures. *Earthquake Engineering & Structural Dynamics*, **29** (9), 1287-1305.

[4] ASCE (2017): Minimum Design Loads for Buildings and Other Structures. *ASCE/SEI 7-16,* American Society of Civil Engineers, Reston, VA.

[5] AISC (2016a): Specification for Structural Steel Buildings. *ANSI/AISC 360-16*, American Institute of Steel Construction, Chicago, IL, USA.

[6] AISC (2016b): Seismic Provisions for Structural Steel Buildings. *ANSI/AISC 341-16*, American Institute of Steel Construction, Chicago, IL, USA.

[7] Zaruma S, Fahnestock LA (2018): Assessment of design parameters influencing seismic collapse performance of buckling-restrained braced frames. *Soil Dynamics and Earthquake Engineering*, **113**, 35-46.

[8] Mathur K., Fahnestock LA, Okazaki T, Parkolap MJ (2011): Impact of residual stresses and initial imperfections on the seismic response of steel moment frames. *Journal of Structural Engineering*, **138** (7), 942-951.

[9] ASCE (2017): Seismic Evaluation and Retrofit of Existing Buildings. *ASCE/SEI 41-17*, American Society of Civil Engineers, Reston, VA.

[10] Harris III JL, Speicher MS (2018): Assessment of Performance-Based Seismic Design Methods in ASCE 41 for New Steel Buildings: Special Moment Frames. *Earthquake Spectra*, **34** (3), 977-999.

[11] Speicher MS, Harris III JL (2016a): Collapse prevention seismic performance assessment of new special concentrically braced frames using ASCE 41. *Engineering Structures*, **126**, 652-666.

[12] Speicher MS, Harris III JL (2016b): Collapse prevention seismic performance assessment of new eccentrically braced frames using ASCE 41. *Engineering Structures*, **117**, 344-357.

[13] Speicher MS, Harris III JL (2018): Collapse Prevention seismic performance assessment of new buckling-restrained braced frames using ASCE 41. *Engineering structures*, **164**, 274-289.

[14] Speicher MS, Dukes JD, Wong KW (2020): Collapse Risk of Steel Special Moment Frames per FEMA P695. *NIST Technical Note 2084*, National Institute of Standards and Technology, Gaithersburg, MD, USA. https://doi.org/10.6028/NIST.TN.2084.

[15] FEMA (2009): Quantification of building seismic performance factors. *FEMA P695*, Federal Emergency Management Agency, Washington DC, USA.

[16] Harris III JL, Speicher MS (2015): Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings Volume 1: Special Moment Frames. *NIST Technical Note 1863-1*, National Institute of Standards and Technology, Gaithersburg, MD, USA. http://dx.doi.org/10.6028/NIST.TN.1863-1.

[17] McKenna F, Fenves GL, Scott MH (2000): Open system for earthquake engineering simulation, version 2.5.0. *University of California, Berkeley*, CA, USA.

[18] Ibarra LF, Medina RA, Krawinkler H (2005): Hysteretic models that incorporate strength and stiffness deterioration. *Earthquake engineering & structural dynamics*, **34** (12), 1489-1511.

[19] Lignos DG, Krawinkler H (2010): Deterioration modeling of steel components in support of collapse prediction of steel moment frames under earthquake loading. *Journal of Structural Engineering*, **137** (11), 1291-1302.

[20] ATC (2017): Guidelines for Nonlinear Structural Analysis for Design of Buildings: Part IIa – Steel Moment Frames. *NIST GCR 17-917-46v2*, Applied Technology Council, Redwood City, CA, USA

[21] Krawinkler H (1978): Shear in beam-column joints in seismic design of steel frames. *Engineering Journal*, **15** (3).

LOADING PROTOCOLS AND BACKBONE CURVES: U.S. AND JAPAN PERSPECTIVES

B. Maison(1), K. Kasai(2), M. Speicher(3)

(1) *Structural Engineer, El Cerrito, USA, maison@netscape.com*
(2) *Specially Appointed Professor, Tokyo Institute of Technology, JAPAN, kasai.k.ac@m.titech.ac.jpx*
(2) *Research Structural Engineer, National Institute of Standards and Technology, USA, matthew.speicher@nist.gov*

Abstract

With the gaining popularity of performance-based engineering, it is essential to have reliable estimates of component (e.g., steel beam or column) performance under actual earthquake loading demands. Physical laboratory tests are traditionally used to determine component inelastic behaviors. A loading protocol refers to the sequence of quasi-static displacements applied to a test specimen to simulate the demands imposed by an earthquake. The vast majority of *standard* tests use loading protocols consisting of fully-reversed cyclic loading with progressively increasing amplitudes. Standard tests are in many cases, unlike the demands posed by earthquakes, and using realistic earthquake loading patterns can lead to different conclusions on component performance. Unfortunately, there are relatively few tests using realistic earthquake loading patterns. Likewise, there are relatively few tests using monotonic loading that may best represent the demands at incipient collapse caused by large near-field type shaking.

In the United States, a component *backbone* curve representing the inelastic force-deformation behavior is typically taken as the envelope of laboratory test hysteresis data. A shortcoming is that the backbone curve is dependent on the loading protocol used in the tests, and customary standard tests can lead to overly pessimistic estimates of component behavior. This can cause rejection of buildings that would otherwise be considered acceptable should component behaviors be based on tests using realistic earthquake loading patterns. It is therefore recommended that future tests consider realistic earthquake loading protocols so that the results are best suited for performance-based engineering.

Alternatively, a Japanese *skeleton* curve concept using decomposition of standard test data may be a way to leverage the current wealth of available standard test data thereby providing a better description of component seismic performance than simply taking an envelope of standard test data. The skeleton curve consists of horizontally stacked hysteretic loops that resemble a monotonic curve. To demonstrate this concept, a simple *adaptive* component model incorporating skeleton curves is presented. The model accounts for both in-cycle and cyclic strength degradation. It is shown to mimic observed monotonic and cyclic behaviors depending on the displacement loading history. The model is simple and can be easily incorporated into structural analysis computer programs used for building evaluation. Ongoing work includes more validation of the adaptive model as well as possible extension to other materials such as reinforced concrete and wood components.

Keywords: loading protocols; performance-based engineering; backbone curves; experimental tests, steel components

Maison, Bruce; Kasai, Kazuhiko; Speicher, Matthew. "Loading Protocols and Backbone Curves: U.S. and Japan Perspectives." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Performance-based engineering, in which a structure is proportioned to meet certain predictable performance requirements, necessitates good estimates of component (e.g., steel beam) inelastic behavior during earthquakes. In the United States, a so-called *backbone* curve is the customary way of describing component behaviors over a range of deformations. It is formulated as an envelope of hysteresis loops from component experimental tests and is a critical factor in component modeling and acceptance criteria [1].

To determine backbone curves, the vast majority of steel component tests use *standard* displacement loading patterns (protocols) that do not necessarily simulate realistic earthquake demands. Results from standard tests can lead to overly conservative component backbone curves (i.e., indicating premature failure), especially for components meeting current ductility requirements in steel construction standards. This can cause rejection of buildings that would otherwise be considered acceptable should component behaviors be based on tests using loading patterns that simulate actual earthquake demands [2, 3].

This paper underscores the influence loading protocols have on backbone curves and demonstrates how a Japanese *skeleton* curve concept can be used in a simple *adaptive* model for steel components. Strength degradation caused by local buckling is separated into in-cycle and cyclic parts. The model incorporates the skeleton curve to mimic observed monotonic and cyclic behaviors depending on the displacement loading history. It better describes inelastic behaviors as opposed to a conventional approach using an envelope of the standard test hysteretic loops. The model can be easily implemented into structural analysis computer programs for the seismic evaluation of structures.

2. Loading Protocols

Figure 1 contrasts a standard protocol to that of a simulated earthquake response from a building undergoing inelastic actions. It consists of a series of fully-reversed displacement cycles having progressively increasing amplitudes. In contrast, the earthquake response has relatively few cycles with a one-direction bias.

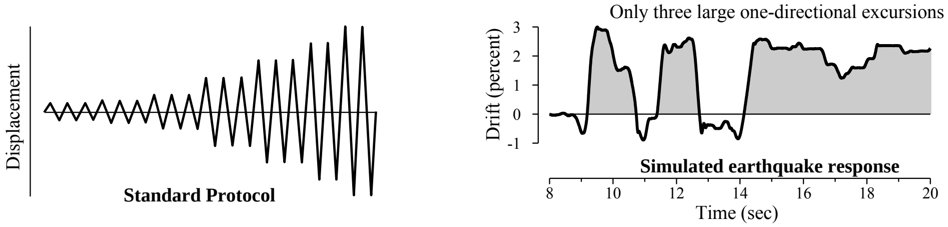


Fig. 1 – Typical standard protocol [4] compared to simulated building inter-story drift earthquake response.

With the advancement and increasing use of nonlinear structural analysis, a better understanding of actual seismic response has led researchers to propose different protocols better reflecting building inelastic seismic response (Figure 2). Such protocols are termed here as *realistic*. Realistic protocols differ from standard protocols, and their use can lead to different conclusions about component performance. Unfortunately, relatively few tests have been conducted using realistic protocols as compared to the plethora of tests using standard protocols. Likewise, relatively few tests have been conducted using monotonic loading that provide insights about component performance at near-collapse displacements [5].

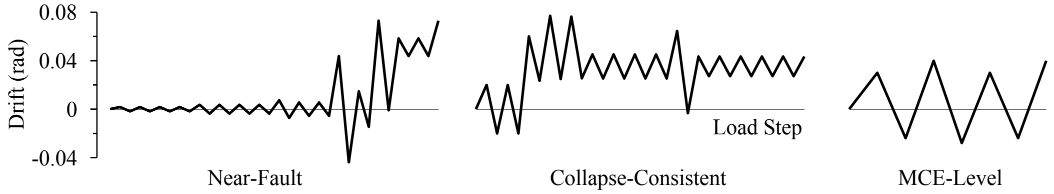


Fig. 2 – Realistic protocols simulating building inelastic drift response during earthquakes: Near-Fault [6], Collapse-Consistent [7], and MCE-Level [8].

3. United States Backbone Curves

Figure 3 shows backbone (envelope) curves from component tests of steel reduced beam section (RBS) connections using standard and realistic loading protocols [9]. The displacement reversal points control where there is an abrupt decline in force giving the impression that the component has a 5 % ultimate drift capacity in the standard test, whereas it has 8 % ultimate drift capacity in the realistic test (1.6-times larger). In reality, the component can have greater capacity than that from either test since the ultimate drifts are an artifact of the protocol reversal points, and not an intrinsic property. This is of particular importance for ductile components such as those designed and constructed in accordance with current steel construction standards such as AISC 341-16 [10]. Nevertheless, realistic tests are most likely better representations of earthquake performance since the loadings mimic actual earthquake demands. Additional shortcomings of standard tests are discussed in reference [8]. Given the abundance of tests conducted using standard loading protocols, it may be advantageous to decompose existing standard test results to gain insights about component performance during earthquakes, as proposed in the next section.

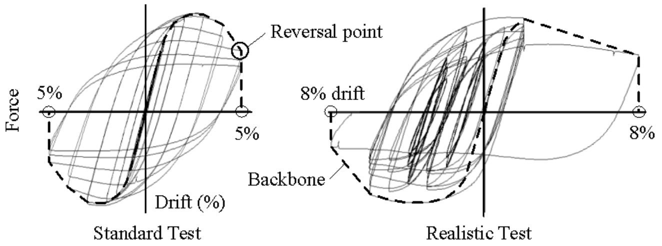


Fig. 3 – Backbone curves from RBS connection tests using standard and realistic loading protocols.

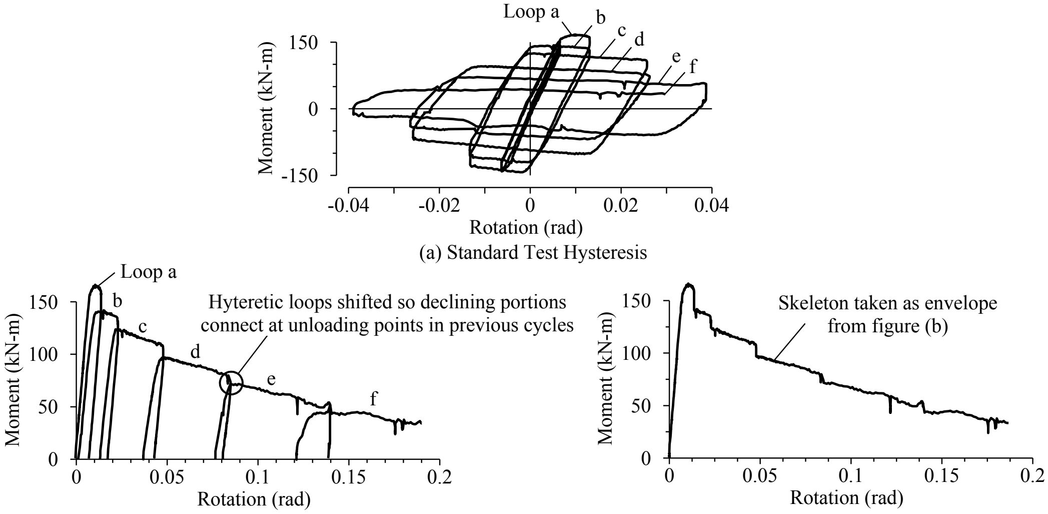
4. Japanese Skeleton Curves

Yamada et al. [11, 12] studied the hysteretic behavior of steel beam-columns having rectangular hollow shaped (RHS) sections. It was found that the cyclic degradation of the moment-rotation relationship caused by local buckling can be represented by a so-called *skeleton* curve. A novel way to decompose the results from a standard test to create a skeleton curve was developed. The process takes the individual hysteretic loops from a standard test and expands them to resemble a monotonic curve. The skeleton curve was then incorporated into an analytical model that reasonably captured RHS inelastic moment-rotation behaviors under random earthquake loadings.

Used here is a modified version of the Yamada skeleton approach as presented by Kimura et al. [13, 14, 15] for steel sections having H-shapes. Figure 4 illustrates the decomposition of the positive-side hysteretic loops to construct a skeleton. The skeleton strength deterioration is based on segments of the hysteretic loops that have declining strength. The segments are shifted horizontally so they connect at the respective unloading rotation in the previous loading cycle. Another skeleton curve can be constructed in a similar manner using the negative-side data.

For the case in Figure 4, the skeleton suggests ductile behavior out to about 0.2 rad that is well beyond the limits of the standard test (0.04 rad). The skeleton curve can serve as an improved (less conservative) backbone curve compared to the conventional approach of taking an envelope of the standard test hysteretic loops. The skeleton curve lacks an abrupt decline in strength like that in conventional backbones (Figure 3) signaling the ultimate rotation. Thus for component acceptance criteria, it may be appropriate to set the ultimate rotation as when the skeleton moment drops below some value (e.g., below 50 % of peak value). However, for computer modeling in building analysis/evaluation, it seems more correct for the component to follow the skeleton even at relatively large rotations.

It should be noted that if a monotonic curve is available, then it can be referenced directly without construction of a skeleton curve. However, there is relatively scant monotonic test data so using decomposition of the plentiful standard test data is an attractive alternative.



(a) Standard Test Hysteresis

(b) Positive-Side Skeleton Construction

(c) Skeleton Curve

Fig. 4 – Skeleton curve derived using positive-side hysteretic loops from standard test.

5. Case Study Using Skeleton Curves

To further explain the concept of skeleton curves and how they can be used in an adaptive model, a set of experimental test data by Kimura et al. [13] for steel H-shaped sections is studied. The test arrangement and hysteretic responses from two tests are shown in Figures 5 and 6, respectively. Case A had no axial force, and Case B had an applied constant compressive axial force of 30 % of the yield force. Conventional backbone (envelope) curves are denoted by red dashed lines in Figure 6.

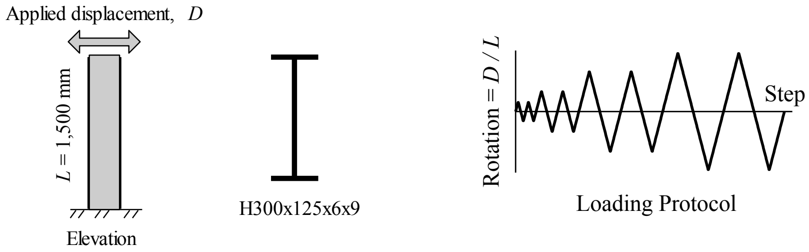


Fig. 5 – Experimental arrangement of standard test.

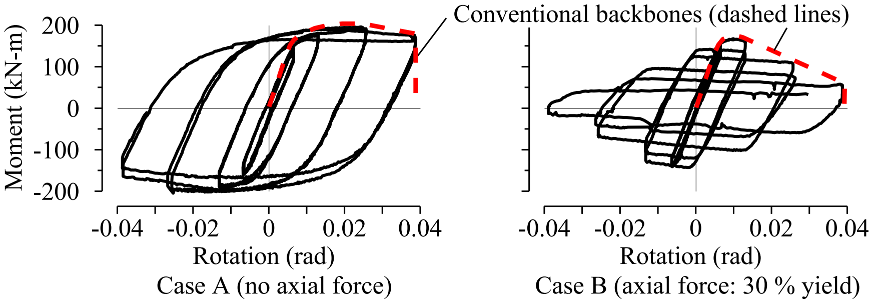


Fig. 6 – Hysteretic response from standard tests.

Figure 7 shows the skeleton curves using the hysteretic loop decomposition technique explained in section 4. The skeleton curves are in reasonable agreement with the monotonic curves for both cases. Case B exhibits earlier strength degradation that is more rapid than that for Case A due to the compressive axial force in Case B promoting local buckling. Conventional backbone curves (red dashed lines), taken as envelopes of the hysteretic loops in Figure 6 indicates ultimate rotation of 0.04 rad defining failure simply due to the fact this was the maximum rotation used in the tests. Both Case A and B would have ultimate rotations well beyond 0.06 rad based on the skeleton curves (Figure 7). Hence, using the skeleton as the backbone curve could have an ultimate rotation for use in acceptance criteria at least 50% greater than the conventional backbones.

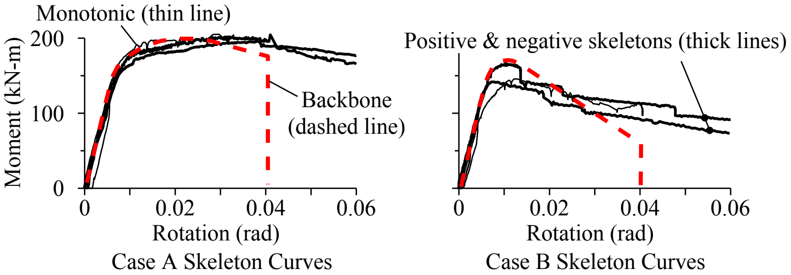


Fig. 7 – Skeleton curves derived from decomposition of standard test hysteretic loops.

Figure 8 shows piecewise linear representations of the skeleton curves from Figure 7. These will be used in a simple adaptive model described in the next section.

Maison, Bruce; Kasai, Kazuhiko; Speicher, Matthew. "Loading Protocols and Backbone Curves: U.S. and Japan Perspectives." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

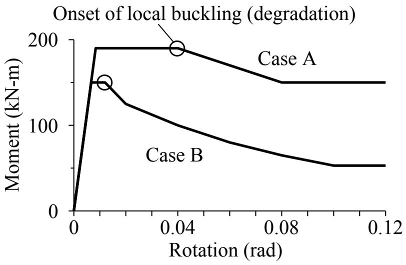


Fig. 8 – Piecewise linear skeleton curves from standard tests.

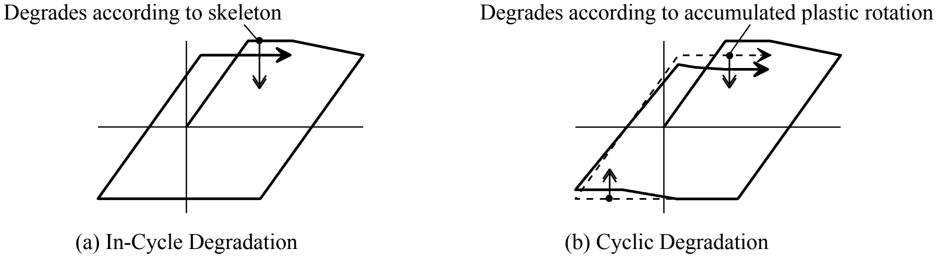
6. Adaptive Model

A simplified version of the Yamada et al. approach for the moment-rotation behavior is presented here. It is based entirely on observations of standard test hysteretic behaviors through use of skeleton curves to account for strength degradation. It should be noted that if a monotonic curve is available, then it can be referenced directly without construction of a skeleton curve. The model is termed *adaptive* because it can reflect monotonic as well as hysteretic strength degradation depending on the deformation loading history.

Strength degradation can be considered as having two parts [16]: in-cycle and cyclic. In-cycle is characterized by loss of strength occurring within a single cycle (e.g., during a monotonic push). Cyclic is delineated by loss of strength occurring in subsequent cycles (e.g., after cycles having the same peak-to-peak displacements).

6.1 In-Cycle Degradation

The skeleton curve serves as the boundary limit for the peak moment. In-cycle degradation is controlled by the skeleton curve as indicated in Figure 9a. The behavior is elasto-plastic and as the rotation increases, the ultimate moment progressively decreases according to the skeleton curve boundary.


Degrades according to skeleton
Degrades according to accumulated plastic rotation

(a) In-Cycle Degradation

(b) Cyclic Degradation

Fig. 9 – Adaptive mode strength degradation controlled by skeleton curve.

6.2 Cyclic Degradation

Cyclic degradation is modeled by a shrinking of the backbone curve according to cyclic actions (Figure 9b). The backbone curve is progressively scaled smaller as a function of the cumulative positive (*APR+*) and negative plastic rotations (*APR-*). The *smaller* of *APR+* or the absolute value of *APR-* is taken as a measure of the cyclic action. A scale factor is computed as the ratio of moments from the skeleton curve (Figure 10). The scaling occurs continuously as the plastic rotations accumulate during the loading history.

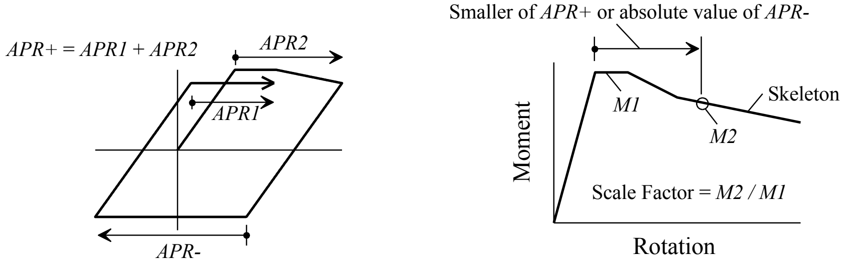
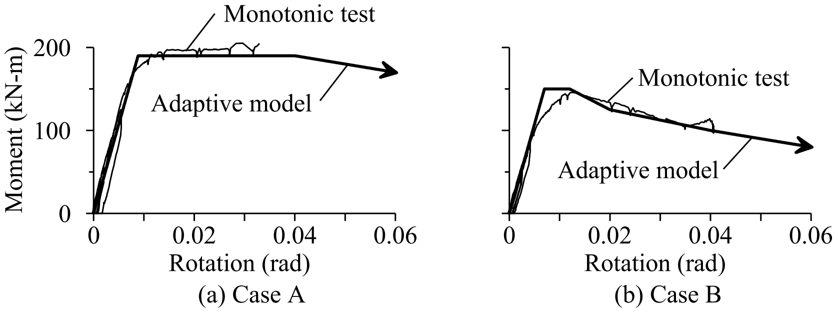


Fig. 10 – Adaptive model cyclic strength degradation feature.

6.3 Comparison to Test Results

The adaptive model was implemented into a computer program and numerical simulations performed. The program has rotation as input and computes the moment according to the adaptive model algorithm described above. A future step is to implement the model into a structural analysis program for building seismic analysis.

Figure 11 compares the results from the adaptive models to monotonic test results. The models produce close fits to the tests. This is expected since there is no cyclic action and therefore no cyclic degradation occurs in the model.

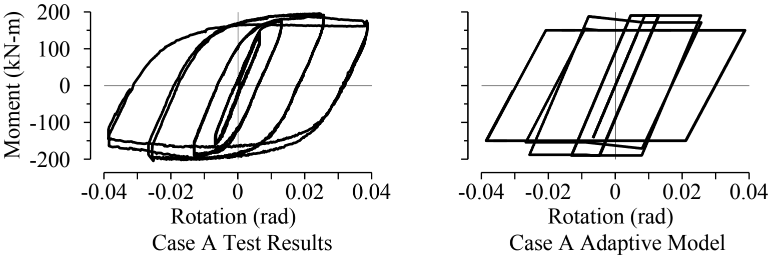


(a) Case A

(b) Case B

Fig. 11 – Comparison of monotonic results.

The Case A adaptive model hysteresis is compared to the standard test in Figure 12. The model exhibits reasonable agreement with the test, albeit the strength degradation is modest.



Case A Test Results

Case A Adaptive Model

Fig. 12 – Comparison of Case A hysteretic response (no axial force).

Figure 13 shows the Case A moment versus cumulative absolute rotation. This can be thought of as a pseudo-time history of moment response. The model reasonably captures the degradation of peak moments in the test and has an excellent fit out to about 0.3 rad cumulative rotation.

Maison, Bruce; Kasai, Kazuhiko; Speicher, Matthew. "Loading Protocols and Backbone Curves: U.S. and Japan Perspectives." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

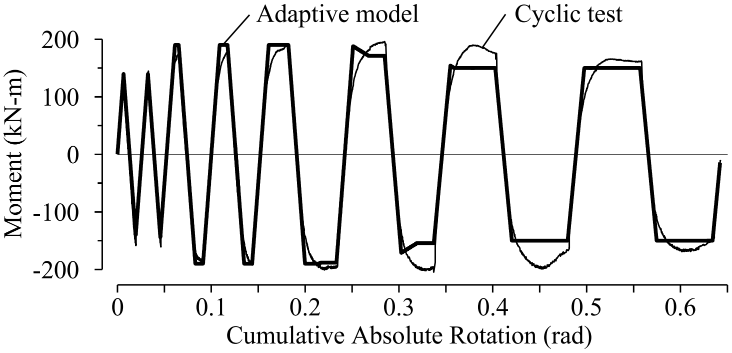
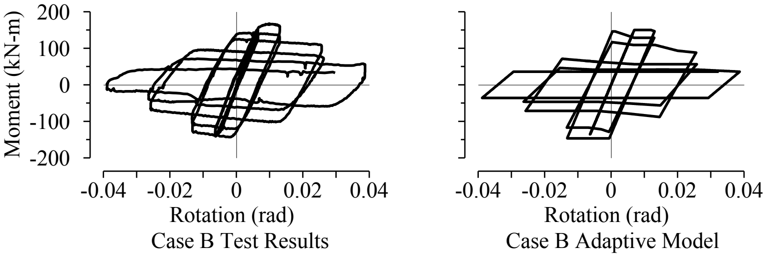


Fig. 13 – Case A (no axial force) moment versus cumulative rotation (pseudo-time history).

The Case B adaptive model hysteresis is compared to the standard test in Figure 14. There is a large amount of strength degradation due to the relatively large axial compression. The model reasonably simulates the test.



Case B Test Results

Case B Adaptive Model

Fig. 14 – Comparison of Case B hysteretic response (axial force: 30 % yield).

Figure 15 shows the Case B moment versus cumulative absolute rotation (pseudo-time history of moment response). The model reasonably tracks the moment degradation in the test.

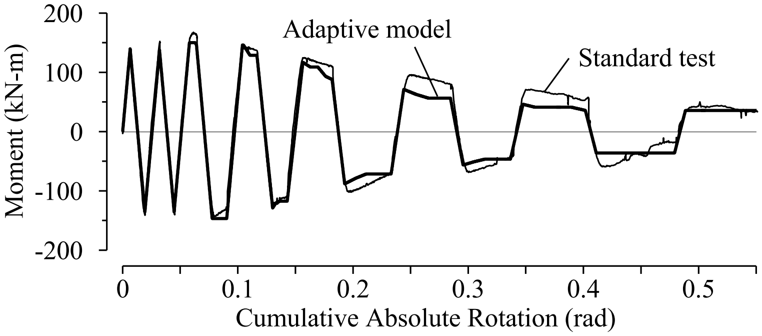


Fig. 15 – Case B (axial force: 30 % yield) moment versus cumulative rotation (pseudo-time history).

The above demonstrates how a simple adaptive model can reasonably mimic component moment response under monotonic as well standard loading patterns. The model incorporates a skeleton curve derived from decomposition of standard test data. Alternatively, should monotonic test curve be available, it can be used in place of the skeleton curve.

7. Realistic Seismic Loading Patterns

The response of the adaptive model from realistic seismic loading patterns (Figure 2) is presented in this section. The near-fault and collapse consistent protocols represent building response at an incipient collapse state. The deformations have a one-direction bias with peak drifts in the 7 to 8 % range. The MCE-Level protocol represents a modern code-conforming building response to a maximum considered earthquake. It also has a one-direction bias but with a smaller peak drift of 4 %.

7.1 Hysteretic Response

Figures 16, 17 and 18 show the adaptive model hysteretic behaviors. Both in-cycle and cyclic degradation are apparent. Case A, having no axial compressive force, has only a modest amount of strength degradation even out to 0.08 rad. This suggests that strength degradation (local buckling) plays only a minor role in the component seismic performance when there is little axial force (e.g., girders in steel moment frame structures). Case B, having significant axial compression, the situation is quite different with significant strength degradation. Hence, columns in steel moment frames are likely to be more susceptible deterioration from the effects of local buckling.

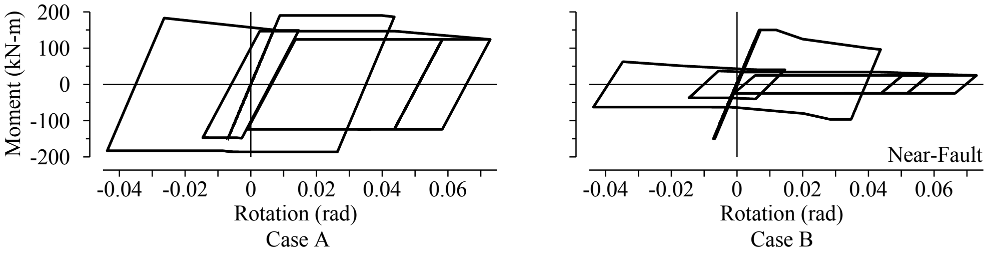


Fig. 16 – Adaptive model hysteresis from *near-fault* loading protocol.

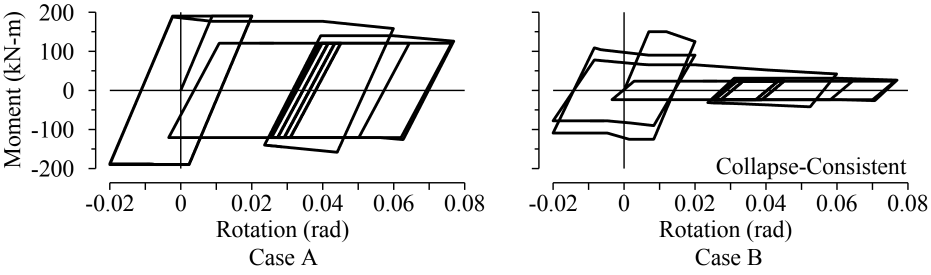


Fig. 17 – Adaptive model hysteresis from *collapse-consistent* loading protocol.

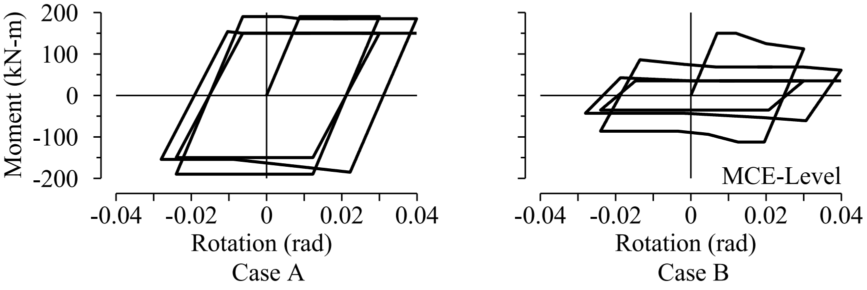


Fig. 18 – Adaptive model hysteresis from *MCE-level* loading protocol.

7.2 In-Cycle vs. Cyclic Degradation

The relative influence of in-cycle and cyclic strength degradation is illustrated in Figures 19 and 20. The adaptive model having both in-cycle and cyclic degradation is compared to the situation having only in-cycle degradation. Case A has a total 34 % strength reduction of which about one-half (17%) is from cyclic degradation. Case B has a huge total strength reduction of 83 % of which about two-thirds (64 %) is from cyclic degradation. For this example, the adaptive model in-cycle and cycle strength reduction features both had significant contributions.

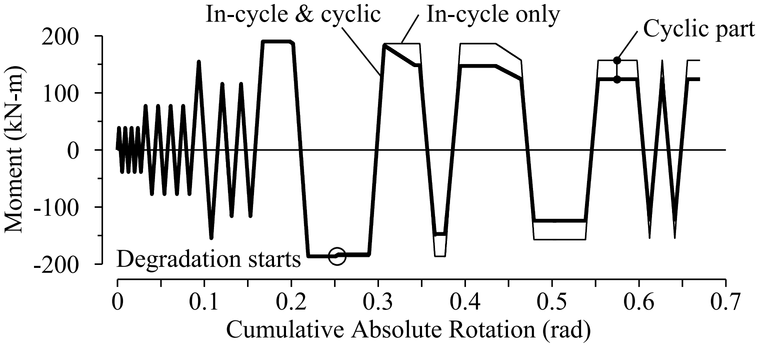


Fig. 19 – Case A (no axial force) moment versus cumulative rotation (pseudo-time history) from *near-fault* loading protocol.

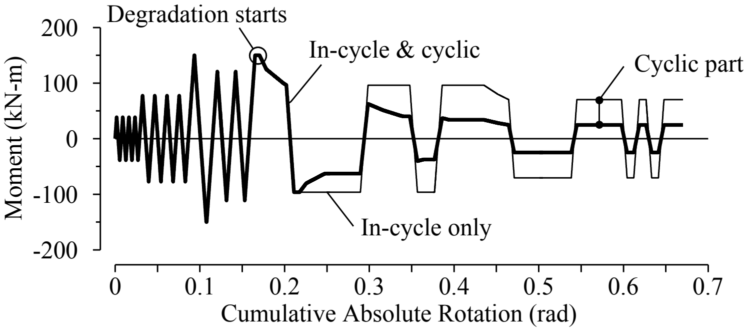


Fig. 20 – Case B (axial force: 30 % yield) moment versus cumulative rotation (pseudo-time history) from *near-fault* loading protocol.

8. Conclusion

Performance-based engineering requires good estimates of component behaviors during actual earthquakes. Customary U.S. practice describes component seismic performance by a backbone curve taken as an envelope of hysteresis loops from experimental tests. Standard tests use loading protocols consisting of fully-reversed cyclic loading unlike the loading patterns posed by earthquakes. Standard tests can lead to overly pessimistic estimates of component behavior.

There are two underlying shortcomings when using standard test data: (1) ductility can be underestimated, which in turn, can lead to very restrictive acceptance criteria, and (2) derived backbone curves used in analysis models for seismic evaluation can lead to over-estimation of peak inelastic displacements, unless those models explicitly account for degradation. These have a compounding effect causing rejection of buildings that would otherwise be considered acceptable should component behaviors be based on tests using realistic earthquake loading patterns. Hence, it is encouraged that future lab tests include realistic earthquake loading protocols so that the results are better suited for performance-based engineering.

There are large numbers of standard tests as opposed to relatively few realistic and monotonic tests. The Japanese skeleton curve concept provides a way to obtain additional useful information from standard test data. Skeleton curves resemble monotonic curves. They can provide less conservative estimates of ultimate rotations (and therefore acceptance criteria) compared to the conventional backbone approach of taking an envelope of the standard test hysteretic loops. In addition, skeleton curves can be used a simple adaptive component model as presented here. The model accounts for both in-cycle and cyclic strength degradation, and is shown to mimic observed monotonic and cyclic behaviors depending on the displacement loading history. The model is simple and can be easily incorporated into structural analysis computer programs used for building evaluation.

Ongoing work includes: more validation of the adaptive model; possible extension to other materials such as reinforced concrete and wood components; and implementation of the adaptive model into structural analysis computer programs to assess the importance of degradation in building global seismic response.

9. Acknowledgement

The authors appreciate Prof. Yamanishi of Hiroshima Institute of Technology, and Dr. Suzuki and Prof. Kimura of Tohoku University in Japan for kindly sharing the experimental test data used in this study. This project is dedicated to the continued cooperation of the Japan and United States earthquake engineering communities. Sharing of our knowledge is the best way toward the advancement of earthquake engineering thereby mitigating hazards posed by earthquakes.

10. References

[1] American Society of Civil Engineers (ASCE), (2017). *Seismic Evaluation and Retrofit of Existing Buildings*, ASCE Standard ASCE/SEI 41-17, Reston, VA.

[2] Speicher MS, Maison BF, (2019). The blind side: using 'canned' loading protocols in seismic testing," *Proceedings of the 12th Canadian Conference on Earthquake Engineering*, Quebec, Canada, June.

[3] Harris, JL, Speicher, MS (2015). *Assessment of First Generation Performance-Based Seismic Design Methods for New Steel Buildings*, Volume 1: Special Moment Frames, National Institute of Standards and Technology (NIST), Gaithersburg, MD. http://dx.doi.org/10.6028/NIST.TN.1863-1.

[4] Applied Technology Council (ATC), (1992). *Guidelines for Seismic Testing of Components of Steel Structures*, ATC-24, Redwood City, CA.

[5] Krawinkler, H (2009). Loading histories for cyclic tests in support of performance assessment of structural components, *Proceedings of the 3rd International Conference on Advances in Experimental Structural Engineering* (3AESE), San Francisco, CA.

[6] Krawinkler H, Parisi F, Ibarra L, Ayoub A, Medina R, (2001). *Development of a Testing Protocol for Woodframe Structures*, CUREE-Caltech Woodframe Project, CUREE Publication W-02.

[7] Yusuke S, Lignos DG, (2014). Development of loading protocols for experimental testing of steel columns subjected to combined high axial load and lateral drift demands near collapse, *Proceedings of the 10th National Conference in Earthquake Engineering*, Earthquake Engineering Research Institute, Anchorage, AK.

[8] Maison BF, Speicher MS, (2016). Loading protocols for ASCE 41 backbone curves, *Earthquake Spectra*, Earthquake Engineering Research Institute, vol. 32, no. 4, November, 2513—2532.

[9] Uang CM, Yu CS, Gilton CS (2000). Effects of loading history on cyclic performance of steel RBS moment connections, *Proceedings of the 12th World Conference on Earthquake Engineering,* New Zealand.

[10] American Institute of Steel Construction (AISC), (2016). *Seismic Provisions for Structural Steel Buildings*, ANSI/AISC 341-16.

[11] Yamada S, Ishida T, Shimada Y (2012). Hysteresis of RHS columns in the deteriorating range governed by local buckling. *Journal of Structural and Construction Engineering*, AIJ, 77(674), 627-636 (in Japanese).

[12] Yamada S, Ishida T, Jiao Y (2018). Hysteretic behavior of RHS columns under random cyclic loading considering local buckling, *International Journal of Steel Structures*, 18(5), 1761-1771.

[13] Kimura Y, Yamanishi T, Kasai K (2013). Cyclic hysteresis behavior and plastic deformation capacity for H-shaped beams on local buckling under compressive and tensile forces, *Journal of Structural and Construction Engineering*, AIJ, 78(689), 1307-1316 (in Japanese).

[14] Suzuki, A, Kimura, Y, Kasai, K (2017). Local buckling behavior and plastic deformation capacity of H-shaped beams under reversed axial forces, *Proceedings of EUROSTEEL Conference*, September 13–15, Copenhagen, Denmark.

[15] Kimura Y, Suzuki A, Kasai K (2019). Estimation of plastic deformation capacity for I-shaped beams with local buckling under compressive and tensile forces, *Japan Architectural Review*, volume 2, 26–41. https://doi.org/10.1002/2475-8876.12066 (translated paper)

[16] Federal Emergency Management Agency (FEMA), (2009). *The Effects of Strength and Stiffness Degradation on Seismic Response*, Technical Report FEMA P-440A, Washington D.C.

Maison, Bruce; Kasai, Kazuhiko; Speicher, Matthew. "Loading Protocols and Backbone Curves: U.S. and Japan Perspectives." Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

Memory update characteristics of carbon nanotube memristors (NRAM$^{®}$) under circuitry-relevant operation conditions

D. Veksler, G. Bersuker, A. W. Bushmaker, M. Mason
MTD, The Aerospace Corporation
Los Angeles CA 90245, USA
dmitry.veksler@aero.org

P. R. Shrestha, K. P. Cheung, J. P. Campbell
NDCD, National Institute of Standards and Technology
Gaithersburg, MD 20899, USA

T. Rueckes, L. Cleveland, H. Luan, D. C. Gilmer
Nantero Inc., 25 Olympia Ave STE B
Woburn, MA 01801, USA

***Abstract*— Carbon nanotubes (CNT) resistance-change memory devices were assessed for neuromorphic applications under high frequency use conditions by employing the ultra-short (100 ps -10 ns) voltage pulse technique. Under properly selected operation conditions, CNTs demonstrate switching characteristics promising for various NN implementations.**

***Index Terms*-- Neuromorphic computing, CNT, RRAM.**

I. Introduction

Mobile neuromorphic computing systems impart specific requirements to memristors for microelectronic "synapses". Carbon nanotube (CNT) nano-switch devices are a promising class of resistive memory technologies due to excellent retention and endurance, and low operating current. A CNT memory cell (Fig. 1) consists of TiN electrodes sandwiching a ~30 nm layer of a conductive fabric representing a disordered network where CNTs are bound to each other via van der Waals forces [1]. Rearrangements of the CNT contacts, Fig. 2, results in higher (~1-10 MΩ) and lower (~10-100 kΩ) cell resistance states. Due to asymmetry of the CNT fabric engineered by varying nanotubes density, orientations, doping, etc., an applied to the bottom electrode positive voltage increases a number of nanotubes connections resulting in higher cell conductance, while negative voltage decreasing conductance of a NRAM device.

To assess the suitability of CNT nano-switch technology for neural network applications, we investigate stability of NRAM responses to program signals under the conditions, which are close to high frequency circuitry operations. For that purpose, an ultra-short (100 ps -10 ns) voltage pulsed technique [2] was employed. Here we focus on deep neural network (DNN) applications where cell resistance (representing synaptic weight) is required to increase/decrease gradually and quasi-linearly depending on the polarity of the programming pulse [3]. To evaluate how synaptic weight is modulated by incoming signals, we investigated dependency of memristor resistance on pulse durations, voltages, and sequences of SET (program operation increasing conductivity) and ReSET (program operation decreasing conductivity) pulses.

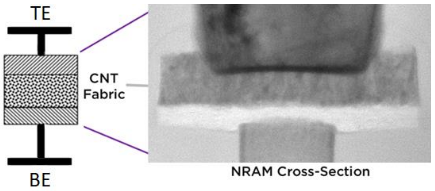


Figure 1. *SEM image of an NRAM cell.*

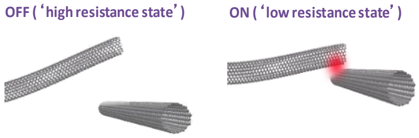


Figure 2. *Schematics of resistance switching between high and low resistance states on a single-point contact.*

II. Experimental setup

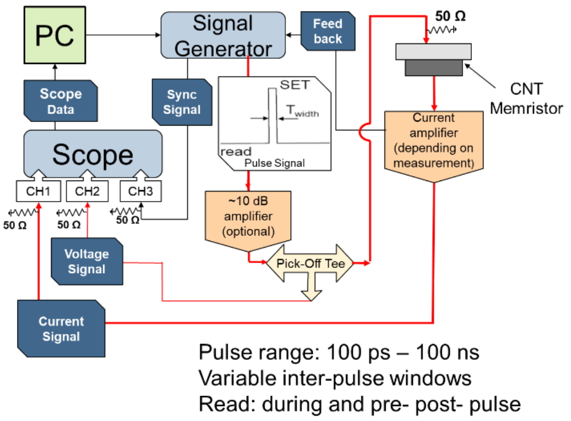
The ultra-short pulse (USP) technique [2,4-8] has several advantages over traditional memristor evaluation methods:

(i) Pulse duration are relevant to actual circuitry operating conditions, and allow realistic assessment of memristor characteristics;

Veksler, Dmitry; bersuker, gennadi; Bushmaker, A; Mason, Maribeth; Shrestha, Pragya; Cheung, Kin (Charles); Campbell, Jason; Rueckes, T; Clevlend, L; Luan, H; Gilmer, D. "Memory update characteristics of carbon nanotube memristors (NRAM) under circuitry-relevant operation conditions." Paper presented at 2020 International Reliability Physics Symposium, Dallas, TX, US. April 28, 2020 - May 30, 2020.

(ii) Short pulses allow precise control of switching energy delivered to devices, which was shown to improve switching stability and reduce variability [2];

(iii) USVP removes the need for a current compliance circuitry because this technique delivers well-controlled small conductance changes.



Pulse range: 100 ps – 100 ns
Variable inter-pulse windows
Read: during and pre- post- pulse

Figure 3. *Experimental setup.*

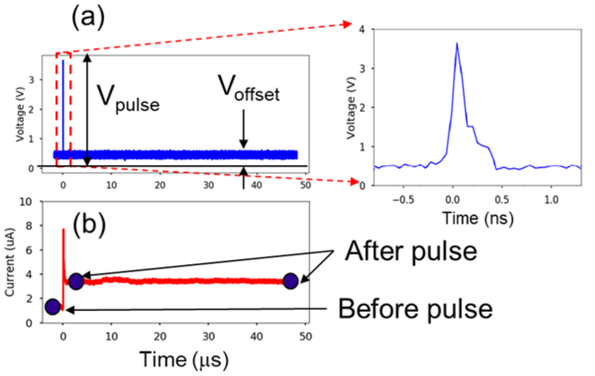


Figure 4. *Applied voltage pulse (a) and resulted current (b) through the NRAM cell, monitored using small DC offset. V_{offset} = 100 – 500 mV.*

Schematic of the experimental setup is shown in Fig. 3. Current through the memristors is monitored using a current amplifier and by applying a constant small DC offset (below the switching voltage threshold) coupled to the programing signals, Fig. 4. Pulses of positive (SET) or negative (ReSET) polarities change conductance to higher and lower values, respectively. Note that CNT memristor devices can operate in a wide conductance range: here we show data on the switching current in the 1 μA and 100 μA range.

III. Results and Discussion

To verify switching stability of NRAM devices, trains of 50 SET pulses followed by 50 ReSET pulses were applied. Pulse duration was fixed at T_{pulse} = 100 ps, and low (1 Hz) repetition rate was used to eliminate cross-correlation (energy-wise) between program pulses. The conductance response to short programing pulses is shown in Fig. 5a,b. both SET and ReSET operations include abrupt and gradual phases of strong and smaller conductance changes, respectively. The first pulse in the 50 SET pulses tends to induce a large conductance increase followed by smaller increases under subsequent pulses; conductance eventually saturates with minimal continuous increase. The current saturation value, $I_{SET,}$ linearly depends on the pulse amplitude, Fig. 6. ReSET typically features an initial gradual conductance decrease followed by an abrupt drop towards the target HRS. In subsequent SET/ReSET switching cycles (performed with sequences of multiple SET and ReSET pulses), stronger initial SETs correlate to larger final steep ReSET drops, Fig. 5c. This indicates that the same sites within the CNT fabric (connections between specific CNTs) control SET-ReSET switching steps: the inter-CNT contact which was activated first (e.g., easier to connect) is the last to be de-activated (harder to disconnect). These sites in the NRAM cell, which are responsible for the initial (SET) - final (ReSET) abrupt conductance changes may represent a cluster of connected CNTs in the film fabric: higher current through the contact in a single CNT pair may assist (via energy dissipation and temperature increase) with bending of surrounding CNTs and promoting contact. Indeed, higher SET voltage results in higher saturation current accompanied by a larger conductance change in both initial SET and final ReSET, Fig. 7 a, b. This is consistent with the proposed understanding that saturation is due to the fact that most (if not all) of available CNT sites capable to connect under a given condition are activated; higher voltage and longer pulse time (i.e. higher emitted energy) increases a pool of activatable sites that leads to both greater initial SET transition and a higher saturation level. To identify operation conditions contributing to conductance saturation in memory updates, we applied program pulses of various durations and variable sequences.

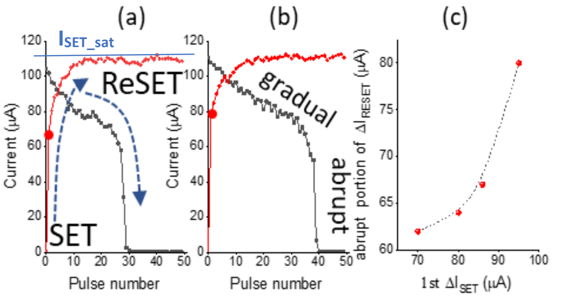


Figure 5. *(a,b) – Example of NMRAM cell response to 50 SET + 50 ReSET pulses in 10th and 20th switching cycles (T_{pulse} = 100 ps). Red circles indicate conductance increase after the 1st SET pulse; (c) correlation between abrupt SET/ReSET phases (multiple cycles on same device): intial SET and final ReSET.*

Veksler, Dmitry; bersuker, gennadi; Bushmaker, A; Mason, Maribeth; Shrestha, Pragya; Cheung, Kin (Charles); Campbell, Jason; Rueckes, T; Clevlend, L; Luan, H; Gilmer, D. "Memory update characteristics of carbon nanotube memristors (NRAM) under circuitry-relevant operation conditions." Paper presented at 2020 International Reliability Physics Symposium, Dallas, TX, US. April 28, 2020 - May 30, 2020.

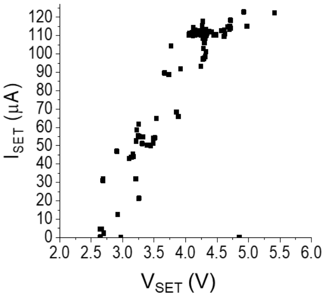


Figure 6. *LRS current after SET operation vs. amplitude of the SET pulse in multiple SET/ReSET cycles (on the same device).*

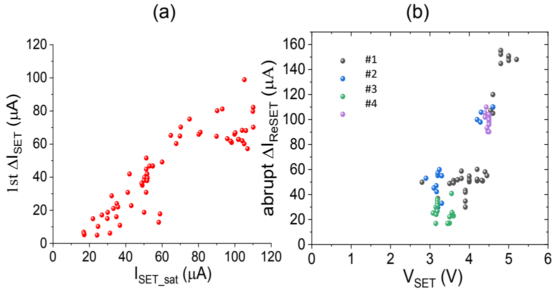


Figure 7. *(a) Correlation between ΔI after 1st SET pulse and saturated conductivity after 50 SET pulses. (b) Abrupt portion of the conductivity update in ReSET operation vs. amplitude of the preceeding train of SET pulses (data for 4 different samples are combined).*

Oscillograms of voltage and current through the NRAM device under 10 ns SET/ReSET pulses are shown in Figure 8. This relatively long voltage pulses of slow rise and fall times allow to directly measure a current during NRAM switching, thanks to low parasitic capacitance of tested devices. Collected data demonstrates strong SET/ReSET asymmetry, Figure 8 c,d. Comparison of switching characteristics of a single SET/ReSET pulse of 10 ns and a train of 5 Set/5 Reset pulses of 2 ns width (separated by 200 ns) further emphasizes the asymmetry between SET and ReSET processes, Fig. 9.

A longer ReSET pulse induces a runaway-like increase of the NRAM cell resistance. In SET, on the other hand, the conductance change is primarily determined by total duration of the applied voltage, either continuous (10 ns pulse) or interrupted (5 x 2 ns).

To assess the possible contribution of temperature increase, SET is performed by applying sequences of short or long pulses separated by a long time-window, Fig. 10. The sequence of 200 ps pulses demonstrates that conductance increase saturates within a short time (covering duration of several initial pulses) contrary to longer 2 ns pulses. This suggests that the energy released during the “short” pulse limits the size of the affected (heated) region of the CNT fabric that reduces possible CNT connections.

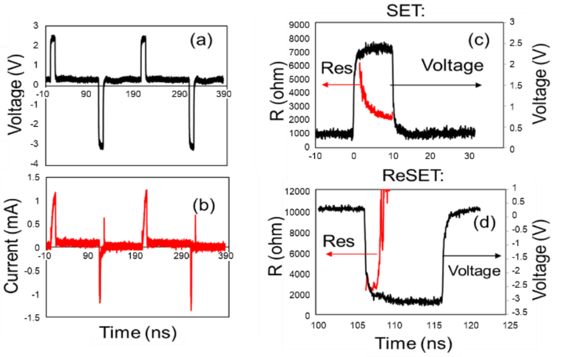


Figure 8. *Oscillograms of voltage (a) and current (b) under 10 ns SET/ReSET pulses; Resistance and voltage evolutions during individual pulses in SET (c) and ReSET (d) operations.*

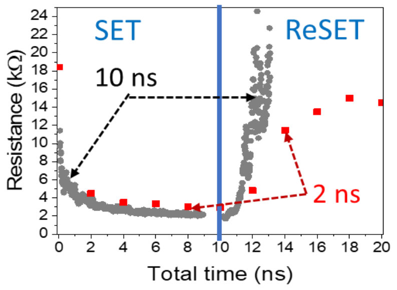


Figure 9. *Comparison of the resistance update in SET and ReSET operations during an applied train of five 2 ns pulses (red dots) and continuous single 10 ns pulse x-axis corresponds to the total time of applied pulse votage. RESET runaway in 10 ns pulse is observed.*

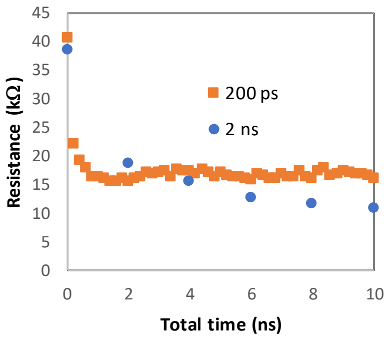


Figure 10. *Comparison of the resistance update in SET operations during applyed trains of five 2 ns pulses (circles) and fifty 200 ps pulses. X-axis corresponds to the total time of the applied pulse voltage.*

This understanding is verified by reducing the time window between the pulses, enabling heat buildup during the pulse sequence. SET cycles are performed by applying repeated pulse sequences of 5 pulses of 200 ps duration; each SET cycle had a specific inter-pulse time window, Fig. 11. The shortest inter-pulse time resulted in higher conductance increase within a given overall pulse cycle duration, while inter-pulse times longer than 200 ps had no effect (consistent with the longer

Veksler, Dmitry; bersuker, gennadi; Bushmaker, A; Mason, Maribeth; Shrestha, Pragya; Cheung, Kin (Charles); Campbell, Jason; Rueckes, T; Clevlend, L; Luan, H; Gilmer, D. ”Memory update characteristics of carbon nanotube memristors (NRAM) under circuitry-relevant operation conditions.” Paper presented at 2020 International Reliability Physics Symposium, Dallas, TX, US. April 28, 2020 - May 30, 2020.

pulses in Fig. 10). This indicates that no heat build-up "around" the forming conductive paths occurred under longer pulses since the emitted energy had sufficient time to dissipate throughout the film. Thus, extremely short pulses (< 200 ps) restrict the SET process by limiting energy supply. This suggests that SET is likely associated with a growing number of CNT contacts gradually establishing a conductive CNT cluster under applied voltage. In ReSET, however, a longer pulse results in drastic resistance increase, Fig. 9, suggesting that a dissociation of CNTs contacts is assisted by the local temperature increase driven by energy released around the conductive paths.

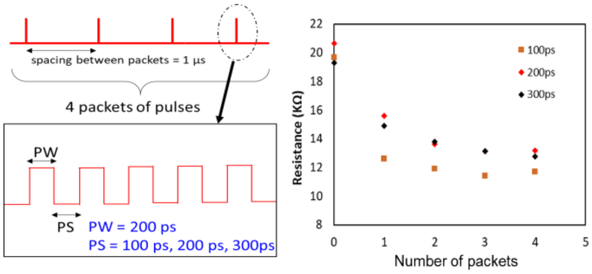


Figure 11. SET: *Resistance updates (r.h.s. graph) under four different pulse sets, each contains five 200 ps pulses. intervals between pulses (PS): 100ps, 200 ps, 300 ps (as marked in l.h.s. graph).*

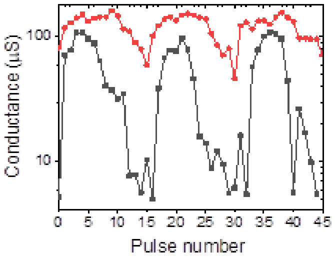


Figure 12. *Conductance switching under 400 ps SET/ReSET pulses Red curve: Vset = 2.5 V and Vreset = -2.4 V; black curve: Vset = 2.3 V Vreset = -2.0 V. . Switching stability and repeatability strongly depends on the pulse amplitude.*

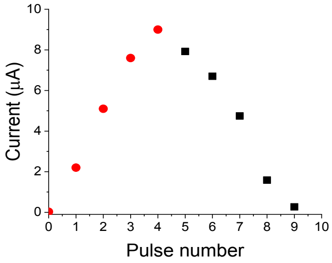


Figure 13. *Quasi-linear and symmetrical gradual conductance update achieved through optimization of switching conditions.*

I. Conclusion

Assessing device operations employing picosecond-range pulses is essential since characteristic time constants in both SET and ReSET processes are found to be on the order of a few hundred picoseconds. NRAM cell shows a stable switching characteristic under sequences of SET/ReSET pulses of 400 ps width at 200 ns intervals, Fig. 12, exhibiting higher resistance under lower pulse voltages. Under properly selected conditions, NRAM demonstrates symmetric quasi-linear memory updates, Fig. 13, making it a promising technology for DNN implementation. Future development of NRAM technology for neuromorphic applications may proceed with the focus on continuous improvements of switching stability and operational current range.

References:

1. D. C. Gilmer, T. Rueckes, L. Cleveland, "NRAM: a disruptive carbon-nanotube resistance-change memory", *Nanotechnology* **29**, p. 134003 (2018).
2. D. Veksler, G. Bersuker, A.W. Bushmaker, P.R. Shrestha, K.P. Cheung, J.P. Campbell, "Switching variability factors in compliance-free metal oxide RRAM", in proc 2019 IEEE International Reliability Physics Symposium, pp. 1-5 (2019).
3. Y. van de Burgt, E. Lubberman, E.J. Fuller, S.T. Keene, G.C. Faria, S. Agarwal, M.J. Marinella, A.A. Talin, A. Salleo "A non-volatile organic electrochemical device as a low-voltage artificial synapse for neuromorphic computing". *Nat. Mater.* **16**, 414–418 (2017).
4. G. Bersuker, D. Gilmer, D. Veksler "Metal oxide resistive random-access memory (RRAM) technology" *Advances in Non-volatile Memory and Storage Technology*, Elsevier (2019).
5. P. R. Shrestha, D. Nminibapiel, J. H. Kim, J. P. Campbell, K. P. Cheung, S.Deora, G. Bersuker, and H. Baumgart, "Energy control paradigm for compliance-free reliable operation of RRAM", in proc. 2014 IEEE International Reliability Physics Symposium, MY.10.1-MY.10.4 (2014).
6. D. M. Nminibapiel, D. Veksler, P. R. Shrestha, Ji-H. Kim, J. P. Campbell. J. T. Ryan, H. Baumgart, and K. P. Cheung, "Characteristics of Resistive Memory Read Fluctuations in Endurance Cycling," *IEEE Electron Device Letters* **38**, pp. 326-329 (2017).
7. G. Bersuker, D. Veksler, D. M, Nminibapiel, P. R Shrestha, J. P Campbell, J. T. Ryan, H. Baumgart, M. S. Mason, K. P. Cheung, *Journal of Comp. Electronics* **16**, pp. 1085-1094 (2017).
8. P.R. Shrestha, D.M. Nminibapiel, J.P. Campbell, J.T. Ryan, D. Veksler, H. Baumgart, K.P. Cheung "Analysis and control of RRAM overshoot current", *IEEE Transactions on Electron Devices* **65**, pp 108-114 (2017).

Veksler, Dmitry; bersuker, gennadi; Bushmaker, A; Mason, Maribeth; Shrestha, Pragya; Cheung, Kin (Charles); Campbell, Jason; Rueckes, T; Clevlend, L; Luan, H; Gilmer, D. "Memory update characteristics of carbon nanotube memristors (NRAM) under circuitry-relevant operation conditions." Paper presented at 2020 International Reliability Physics Symposium, Dallas, TX, US. April 28, 2020 - May 30, 2020.

Simplified Ray Tracing for the Millimeter Wave Channel: A Performance Evaluation

Mattia Lecci*, Paolo Testolina*, Marco Giordani*, Michele Polese*,
Tanguy Ropitault†, Camillo Gentile†, Neeraj Varshney†, Michele Zorzi*

*Department of Information Engineering, University of Padova, Italy, email:{name.surname}@dei.unipd.it
†National Institute of Standards and Technology (NIST), Gaithersburg, MD, 20899 USA, email:{name.surname}@nist.gov

***Abstract*—Millimeter-wave (mmWave) communication is one of the cornerstone innovations of fifth-generation (5G) wireless networks, thanks to the massive bandwidth available in these frequency bands. To correctly assess the performance of such systems, however, it is fundamental to have reliable channel models, based on a deep understanding of the propagation characteristics of the mmWave signal. In this respect, ray tracers can provide high accuracy, at the expense of a significant computational complexity, which limits the scalability of simulations. To address this issue, in this paper we present possible simplifications that can reduce the complexity of ray tracing in the mmWave environment, without significantly affecting the accuracy of the model. We evaluate the effect of such simplifications on link-level metrics, testing different configuration parameters and propagation scenarios.**

I. Introduction

The next generation of Cellular and Wireless Local Area Network (WLAN) will be the first to exploit millimeter wave (mmWave) frequencies to provide connectivity in the access network, i.e., in the links between base stations and mobile users. In particular, 3rd Generation Partnership Project (3GPP) NR has been designed to support a carrier frequency up to 52.6 GHz in Release 15 [1], and future Releases will consider an extension to 71 GHz and the sub-THz band [2]. Similarly, IEEE 802.11ad and 802.11ay exploit the unlicensed bands at 60 GHz [3]. The mmWave frequencies, indeed, feature large chunks of untapped bandwidth that can increase the data rate provided to the end users, making it possible to target the 5th generation (5G) requirements of ultra-high peak throughput (20 Gbps) and average user experienced rate (50-100 Mbps) [4]. Moreover, the small wavelength enables the design of antenna arrays with tens of elements in a small form factor, which could fit even smartphones or VR headsets.

The propagation characteristics of the Radio Frequency (RF) signals in these frequency bands, however, complicate the design of reliable communication systems [5]. First, the high propagation loss, which is proportional to the square of the carrier frequency, limits the coverage region of the mmWave base stations. This can be compensated by using beamforming with large antenna arrays, which could concentrate the power in narrow, directional beams and increase the link budget. Additionally, mmWave signals can be easily blocked by obstacles (e.g., vehicles, buildings, human bodies), preventing direct Line-of-Sight (LoS) communications. Furthermore, at mmWave frequencies, the increased diffraction results in deep shadow regions, thus further degrading propagation performance [6]. By considering the combination of these phenomena, the mmWave channel appears extremely volatile to mobile users, whose quality of experience may be poor unless a proper network design is adopted.

A. Channel Modeling for mmWaves

As experimental platforms and testbeds at mmWaves are still at an early development stage [7], [8], analysis and simulation play a fundamental role for the performance evaluation of novel solutions for mmWave networks. Given the aforementioned behavior of the channel at such high frequencies, and the interplay with network deployment choices and beamforming design, the accuracy of analysis and simulation depends on that of the channel model even more than at conventional sub-6 GHz frequencies. Therefore, the research community has developed a number of channel modeling tools for mmWaves, with a varying degree of complexity and accuracy. Stochastic and analytical models are based on the combination of random variables fitted on traces and measurements, and are widely used for analysis [9], [10] and system-level simulations [11], [12]. However, the generality of these models and their stochastic nature fits poorly with the need to accurately characterize specific scenarios. Additionally, most stochastic models may not properly characterize the specific features of the mmWave channel that may affect the overall system performance, such as the temporally- and spatially-consistent updates of the LoS condition and the evolution of each single Multi Path Component (MPC).

These modeling challenges are instead addressed by Ray Tracers (RTs), which have been used to precisely characterize the propagation of RF signals in specific scenarios [13], [14], [15]. With ray tracing, the channel is modeled in terms of MPCs that generate from a certain location and angle of departure, are reflected (and, in complete models, diffused) on the scattering surfaces of the scenarios, and reach the position of the receiver with a certain angle of arrival, delay and power [16]. As the generation of MPCs is purely based on the geometry of the scenario, the channel is as accurate as the description of the scenario, and the MPCs are consistent

This work was partially supported by NIST under Award No. 70NANB18H273. Mattia Lecci and Paolo Testolina's activities were supported by *Fondazione CaRiPaRo* under the grants "Dottorati di Ricerca" 2018 and 2019, respectively.

Lecci, Mattia; Testolina, Paolo; Giordani, Marco; Polese, Michele; Ropitault, Tanguy; Gentile, Camillo; Varshney, Neeraj; Zorzi, Michele. "Simplified Ray Tracing for the Millimeter Wave Channel: A Performance Evaluation." Paper presented at 2020 Information Theory and Applications Workshop, San Diego, CA, US. February 02, 2020 - February 07, 2020.

with the mobility model of the communication endpoints. Additionally, ray tracers can be easily integrated into system-level simulators, by computing the channel matrix $\mathbf{H}$ that combines the different multipath components and the antenna arrays of the network nodes.

With currently available channel modeling tools, however, a higher accuracy translates into a higher computational complexity for the MPC generation and the simulations. As we discuss in [17], the complexity is proportional to two elements, i.e., the number of MPCs which need to be combined to generate the channel matrix $\mathbf{H}$, and the number of antennas at the two endpoints of the communication link (which represents the number of columns and rows of $\mathbf{H}$).

B. Contributions

Based on the above introduction, in this paper we investigate whether it is possible to improve the trade off between accuracy and complexity in mmWave simulations, by studying simplification techniques for ray tracers that speed up the simulations and the ray tracer itself. Specifically, we consider processing only MPCs whose received power is above a certain threshold, which is relative to the strongest MPC, and limit the maximum number of reflections for each MPC. Our results show that it is possible to decrease the complexity of the simulations with a minimal reduction in accuracy, with respect to the baseline ray tracer implementation (i.e., without simplifications).

These promising results are a first step towards understanding and isolating which are the most fundamental channel modeling components at mmWave frequencies, and could stimulate further investigations into whether it is possible to develop simplified channel models (e.g., to be used also for mathematical analysis) that are more representative of the mmWave propagation than the widely-used Nakagami-m or Rayleigh fading models [18].

The rest of the paper is organized as follows. In Section II we provide details on the ray tracer we consider as baseline. We then introduce possible simplifications in Section III, while performance results are discussed in Section IV. Finally, we conclude the paper in Section V.

II. The Millimeter Wave Ray Tracer

To simulate a realistic channel, an open-source MATLAB ray tracer was used[1]. The ray tracer was built with mmWave propagation in mind and for this reason, given the deep shadow effect that diffraction yields at such high frequencies [6], only specular reflections are considered. In this work, diffuse reflections are ignored, but their importance is undoubted and will thus be part of our future analysis. Currently, the ray tracer accepts Computer-aided Design (CAD) files in AMF format with scenarios described by triangles of different materials.

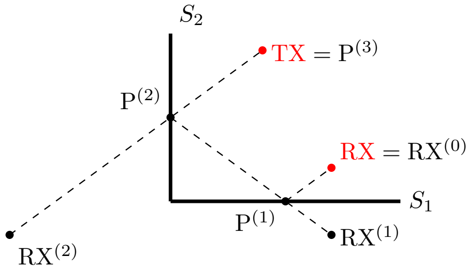


Fig. 1: Visualization for the Method of Images algorithm for a second-order reflection ($N = 2$).

Specular reflections are computed using the Method of Images (MoI), a basic principle from antenna theory [19]. Given two points in 3D space, i.e., the Transmitter (TX) and the Receiver (RX), and a surface S, the MoI defines the virtual image of the RX ($\mathrm{RX}^{(1)}$) to be the reflection of the RX ($\mathrm{RX}^{(0)}$) across the given surface S. By joining the TX with the $\mathrm{RX}^{(1)}$, it is possible to easily compute the point of specular reflection $\mathrm{P}^{(1)}$ as the intersection of the segment with S. It is necessary, though, to check if the reflection point is inside the bounded surface, otherwise the reflection will be discarded. Furthermore, it is also necessary to check that every other surface of the scenario does not intersect the two segments at any point, otherwise the whole ray will be considered obstructed and thus discarded.

When multiple reflections are considered, the MoI applies recursively. Specifically, given an array of reflecting surfaces $\mathcal{S} = (S_1, \dots, S_N)$, $\mathrm{RX}^{(n)}$ is computed as the virtual image of $\mathrm{RX}^{(n-1)}$ for surface S_n, $n = 1, \dots, N$. Then, defining $\mathrm{P}^{(\mathrm{N+1})} = \mathrm{TX}$, the reflection point $\mathrm{P}^{(\mathrm{n})}$ is computed as the intersection between the surface S_n and the segment joining $\mathrm{P}^{(n+1)}$ and $\mathrm{RX}^{(n)}$. Finally, a check is needed on every path segment $\left(\mathrm{P}^{(n)}, \mathrm{P}^{(n+1)}\right)$ to asses whether it is obstructed by any triangle of the environment. Fig. 1 shows a visual example of the MoI algorithm.

To compute all possible reflections between the RX and TX, a *reflection tree* is created, based on the geometric information extracted from the CAD file. In a reflection tree, all nodes except the root (the TX) correspond to triangles of the environment. For each node, its children coincide to all the visible triangles of the environment with respect to that node. Thus, the depth of the tree corresponds to the maximum reflection order $\eta_{\max}$ (given as an input configuration parameter), i.e., the maximum number of reflections per MPC that the RT computes, and each path from the root to a node at depth d corresponds to an ordered array of d reflecting surfaces. By following all the paths for each tree depth d, all possible array of triangles are tested and thus every reflected ray is computed.

An accurate profiling of the software shows that the most demanding parts of the RT operations are the geometric computations (i.e., computing the position of virtual RXs, computing the point of specular reflection, check if the point is inside the bounded surface) and the obstruction checks. The complexity of the geometric computations is proportional to $\eta_{\max}$, while obstruction checks scale both with $\eta_{\max}$ (every segment has to be checked) and the environment complexity (any triangle can potentially block the propagation of the ray).

Finally, if the ray reflections are valid and it is not ob-

[1] Ray tracer implementation: https://github.com/wigig-tools/qd-realization

structed, path gain is computed as

$$PG = \left(\frac{\lambda}{4\pi d}\right)^2 - \sum_{n=1}^{N} RL_n, \qquad (1)$$

where λ is the wavelength (that is a function of the carrier frequency f_c), d is the total distance traveled by the ray, and RL_n is the reflection loss of the material associated to the n-th reflecting surface [16]. Together with the delay, phase, angle of departure, and angle of arrival, the path gain is returned as an output by the ray tracer and written to a file in a specified format, which can be fed as input to other simulators (e.g., link-level or system-level simulators) to compute the channel between the two nodes.

To further simplify the software, no polarization is considered, and rays reflected by a surface experience a 180° phase rotation and a reflection loss of $RL_n = 7$–25 dB, depending on the material but irrespective of the angle of incidence.

III. How to Simplify the Millimeter Wave Channel

As introduced in Sec. II, the number of multipath components of the channel between two endpoints heavily affects both the RT computational complexity and the performance of network-level simulators [17]. For this reason, in this work we propose two different strategies to reduce the total number of MPCs, and analyze the effects that this simplification yields on the system behavior. The two strategies we introduce are based on considerations related to the power of each single MPC, based on the idea that weak MPCs do not significantly contribute to the overall signal at the receiver.

For the first simplification approach, we reduce the number of reflections $\eta_{\max}$ that the RT computes. Indeed, when considering an increasing number of reflections, the MPC experiences a decreasing path gain PG, as the absorption on the reflecting surfaces severely degrades the power of the ray and the length of the ray increases. Therefore, MPCs that bounce across multiple scattering surfaces do not impact very much the power at the receiver, and can be omitted from the RT computations. Limiting the maximum reflection order corresponds to setting a bound to the depth of the reflection tree, whose size, as mentioned in the previous section, is exponential on the maximum reflections order $\eta_{\max}$. Therefore, reducing $\eta_{\max}$ reduces the complexity of both the RT and of the network simulators that use it to model the channel.

Following similar considerations, the second strategy aims at discarding the weakest MPCs based on how low their path gain is, regardlessly of how many reflections they actually experience. Using this method, the path gain PG for a ray still needs to be computed, therefore the geometric operations will not be spared. However, the obstruction check may not need to be performed in the case PG is below a certain threshold, and the ray is not accounted for when computing the channel matrix $\mathbf{H}$. The method we propose is based on a threshold γ_{th} which is relative to the $PG_{\max}$ of the strongest MPC for a given channel realization. Notably, the rays with path gain $PG/PG_{\max} < \gamma_{\text{th}}$ are discarded. The path gain associated with the strongest ray $PG_{\max}$ is updated on-line.

TABLE I: Parameters used for simulations.

P_{TX}	30 dBm	TX Array Config.	8×8
f_c	60 GHz	RX Array Config.	4×4
Noise Figure (F)	5 dB	Element pattern	Omni-directional
Bandwidth	400 MHz	Element spacing	$\lambda/2$

Selecting the MPCs with a relative, rather than an absolute threshold, makes it possible to dynamically adapt the simplification to the actual quality of the channel. For example, in Non-Line-of-Sight (NLoS) conditions, the strongest MPC will be given by a reflected ray. Therefore, its path gain $PG_{\max}$ will be comparable to that of a higher number of MPCs (given by other reflections) than in a LoS scenario, where the strongest ray is the direct path, with a much higher power than the reflections. If the reflections have a PG similar to $PG_{\max}$, then the receiver experiences a strong fading. In this case, a *relative* threshold combines accuracy (in NLoS, significant MPCs are still computed) and reduction in complexity (in LoS, several negligible reflected MPCs are not accounted for), more than an absolute threshold, which would apply the same cut in both cases.

We implemented the proposed simplification techniques on top of the open-source ray tracer described in Sec. II. Although such methods are beneficial from a computational point of view, they may have some drawbacks, depending on the reliability requirements of the application for which the ray tracer is used. The most evident downside is that the power spatial distribution can be affected in complex and non-foreseeable ways, as we will show in Sec. IV.

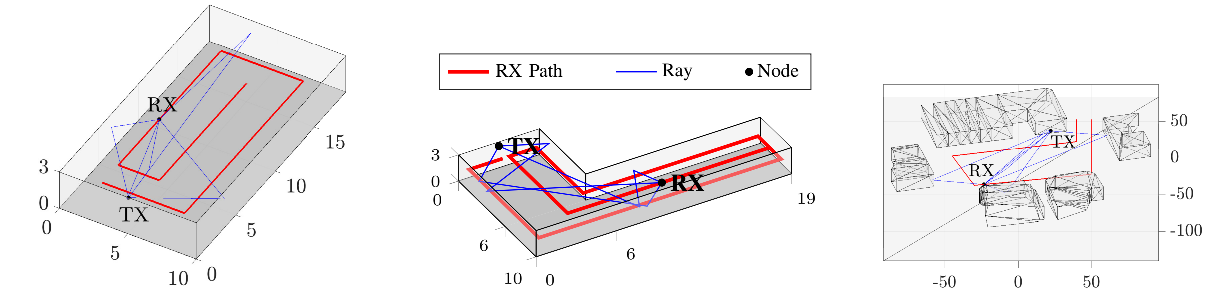
IV. Performance Results

In this section we evaluate through simulations the effects of the ray tracer simplifications we presented in Sec. III. In particular, we investigate the impact of (i) the maximum number of reflections $\eta_{\max}$ per MPC, and (ii) the received power threshold γ_{th} (relative to the strongest path) below which MPCs are are discarded by the ray tracer.

Three scenarios are defined as follows:

1) *Indoor1*: A simple scenario of a box-like room (Fig. 2a) of size $10 \times 19 \times 3$ m. A TX is positioned close to the ceiling, half-way through the wall at $(5, 0.1, 2.9)$ m while the RX, 1.5 m tall, moves inside the room at a speed of 1.2 m/s in spiral-like motion;
2) *L-Room*: Details of this scenario are shown in Fig. 2b. Similarly to the *Indoor1* scenario, the RX of the same height as the previous one moves at a speed of 1.2 m/s;
3) *Parking-Lot*: Outdoor scenario with building around a parking area of about 120×70 m. A TX is positioned on a building 3 m high and a RX is moving at a speed of 4.17 m/s (15 km/h) around the parking lot (see Fig. 2c).

All scenarios have been sampled every 5 ms, thus a total of approximately 9 000, 12 500, and 15 000 time-steps respectively. A list of parameters used in our simulations is shown in Table I. Optimal single-stream SVD-based beamforming is used.



(a) *Indoor1*

(b) *L-room*

(c) *Parking Lot*

Fig. 2: Visual representations of the proposed scenarios

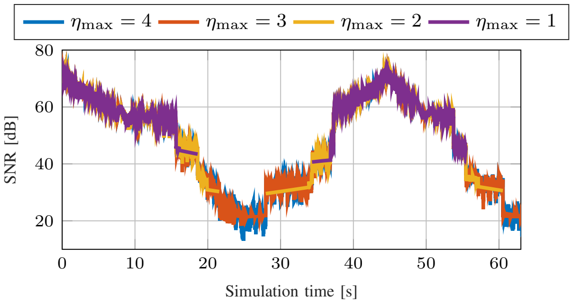


Fig. 3: Temporal evolution of the SNR experienced when a test RX moves in the *L-room* scenario vs. $\eta_{\max}$, fixing $\gamma_{\text{th}} = -\infty$.

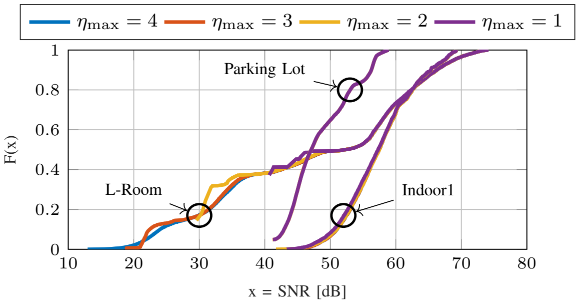


Fig. 4: Cumulative Distribution Function of the SNR when the test RX moves in different simulation scenarios vs. $\eta_{\max}$, with $\gamma_{\text{th}} = -\infty$.

The following performance metrics are considered:[2]

- The RT simulation time T_{RT} [s], i.e., the time taken by the RT software to compute the channel between each pair of nodes at each time-step;
- The MATLAB Net. Sim. Time T_{run} [s], i.e., the time taken by our custom MATLAB simulator to compute the relevant metrics starting from the output of the RT software;
- The Normalized Root Mean Square Error (NRMSE) of the Signal-to-Noise Ratio (SNR), an accuracy indicator that compares the SNR Γ_t experienced when the most accurate RT settings (e.g., with $\eta_{\max} = 4$ and $\gamma_{\text{th}} = -\infty$ for the *L-room* scenario) are considered and the SNR $\hat{\Gamma}$ experienced when different combinations of RT simplifications are applied. Formally, if σ_Γ represents the standard deviation of the baseline SNR Γ, we have

$$\text{NRMSE} = \frac{\text{RMSE}}{\sigma_\Gamma} = \frac{\sqrt{\mathbb{E}\left[\left(\Gamma - \hat{\Gamma}\right)^2\right]}}{\sigma_\Gamma}. \quad (2)$$

In Fig. 3 we consider two nodes and measure the SNR that the probing RX experiences when moving along the path shown in Fig. 2b vs. the maximum number of reflections per MPC $\eta_{\max}$. Rapid variations in the SNR are due to MPCs interfering constructively and destructively, since they observe slightly different path lengths and at least 5 of them have similar path gain in the LoS regions, specifically, the LoS ray and the first-order reflections from ceiling, floor and the two side walls, showing the strong fading in the order of $\lambda = 5$ mm (at 60 GHz). First, we notice that the SNR evolves consistently with the mobility of the RX: the SNR suddenly degrades when the RX enters a NLoS condition and is maximized when it is in LoS with its serving TX, i.e., around time $t = 0$ s and $t = 45$ s. At first glance, it appears that the effect of the RT simplifications is not negligible. In particular, considering the lowest possible value of the relative threshold, i.e., $\gamma_{\text{th}} = -\infty$, the trend of the SNR visibly changes when progressively limiting the maximum number of reflections for each MPC. The impact of those simplifications is particularly evident when the RX operates in NLoS, i.e., when the number of MPCs as little as one (when fading stops) or even none (when no power is received). Despite the above considerations, in the following results we will show more explicitly the accuracy vs. speed trade-off of these parameters in the different scenarios. Furthermore, we will suggest working points for which computation time is significantly reduced with only minor effects on the accuracy of the model.

In Fig. 4 we plot the Cumulative Distribution Function (CDF) of the SNR experienced in different scenarios as a function of the parameter $\eta_{\max}$. The abrupt termination of the CDFs for the *L-Room* and *Parking Lot* scenarios is due to positions of RX/TX for which no ray was able to reach

[2] In this paper, we focus on low-layer performance metrics. In turn, investigating the impact of the proposed RT simplifications on higher-layer performance metrics represents a very interesting research topic that will be part of our future work.

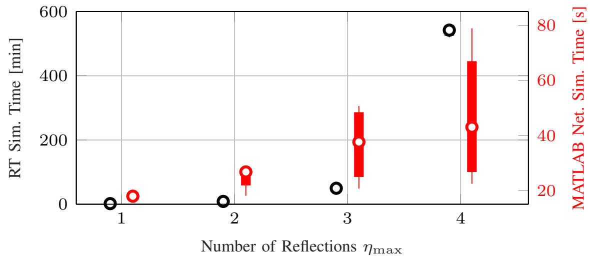


Fig. 5: Box-plots representing the computation time vs. $\eta_{\max}$ when a test RX moves in the *L-room* scenario for all values of γ_{th}. Each box is delimited by the first and the third quartiles of the simulation time, the box's center dot represents the median of the simulation time, and the lines extending from the box (*whiskers*) indicate variability outside the upper and lower quartiles. Each box includes every combination of γ_{th}.

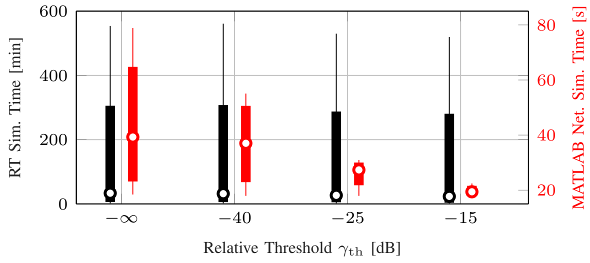


Fig. 6: Box-plots of the computation time vs. γ_{th} when a test RX moves in the *L-room* scenario for all values of $\eta_{\max}$. Each box is delimited by the first and the third quartiles of the simulation time, the box's center dot represents the median of the simulation time, and the lines extending from the box (*whiskers*) indicate variability outside the upper and lower quartiles. Each box includes every combination of $\eta_{\max}$.

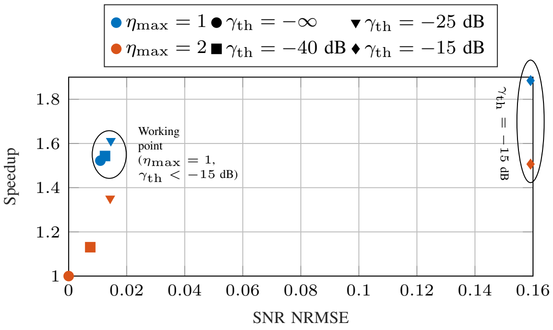


Fig. 7: Speedup vs. SNR NRMSE for different combinations of the RT simplifications when a test RX moves in the *Parking Lot* scenario. $N_{\text{run}} = 1000$ simulations are considered.

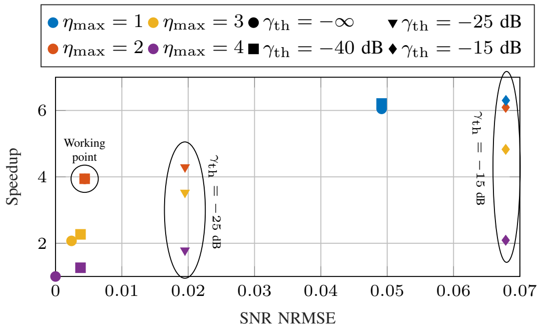


Fig. 8: Speedup vs. SNR NRMSE for different combinations of the RT simplifications when a test RX moves in the *L-room* scenario. $N_{\text{run}} = 1000$ simulations are considered.

with the given $\eta_{\max}$, thus resulting in a complete outage. We observe that, unlike in the *L-room* scenario, in the *Parking Lot* and *Indoor1* scenarios the RX preserves the LoS with its serving TX for the whole duration of the simulation, thereby maintaining very high values of SNR, i.e., above 40 dB. Moreover, Fig. 4 shows that, while in the LoS regime it is possible to reduce the number of reflections $\eta_{\max}$ for each MPC with a minor impact on the accuracy, in the NLoS regime of the *L-room* scenario (i.e., the leftmost part of the figure) the same operation significantly reshapes the CDF of the SNR, thereby confirming the results we obtained in Fig. 3.

On the other hand, decreasing $\eta_{\max}$ speeds up the simulations by several orders of magnitude, as exemplified by the boxplots in Figs. 5 and 6. We can see that the MATLAB simulation time can be reduced by a factor up to 2.4× going from $\eta_{\max} = 4$ to $\eta_{\max} = 1$. The improvement is even more remarkable considering the RT simulation time: the speedup is as significant as 25× considering $\eta_{\max} = 3$, and even 275× for $\eta_{\max} = 4$. Fig. 5 also shows that the configurations with $\eta_{\max} = 3, 4$ exhibit very diverse simulation run time, which is an indication of the increased variability of the channel due to scattering and reflection of the MPCs from nearby surfaces. Similarly, the box-plot in Fig. 6 illustrates that the speedup factor is inversely proportional to the relative threshold γ_{th}, since higher values of γ_{th} makes it possible to reduce the number of MPCs to be processed by the ray tracer as described in Section III, which represents one of the most computationally intensive steps of a simulation and which directly impacts on the channel generation process.

Overall, it is possible to identify which level of simplification is most adequate, i.e., which one provides accurate results while minimizing the overall simulation time. To this aim, in Figs. 7 and 8 we plot the trade-off between simulation speedup and NRMSE of the SNR for the outdoor *Parking Lot* and *L-room* scenarios, respectively. Since, typically, simulations are used to evaluate how changing a set of parameters affects the network performance, and should be repeated with several random seeds to increase the robustness of the obtained results, simulation campaigns would reuse the same RT channel traces for hundreds or thousands of simulations. For this reason, we consider the speedup relative to the total campaign time, which is roughly equal to $T_{\text{TOT}} = T_{\text{RT}} + N_{\text{runs}} T_{\text{run}}$, where T_{RT} is the RT computation time, N_{runs} is the number of independent simulations that are run, and T_{run} is the network simulation run time.

For the *Parking Lot* case (Fig. 7), we can see that all the investigated combinations of simplifications with $\gamma_{\text{th}} < -15$ dB deliver similar values of SNR NRMSE, while reducing the computational complexity with respect to the baseline implementation (i.e., with $\eta_{\max} = 2$ and $\gamma_{\text{th}} = -\infty$). In this scenario, the measured power has limited contribution from the reflected rays, and the optimal approach would be to opt for the configuration with $\eta_{\max} = 1$ and $\gamma_{\text{th}} = -25$ dB: the corresponding speedup is around 60% compared to the baseline ray tracing model.

For the *L-room* case (Fig. 8), instead, it is possible to identify two operational regimes. On the one hand, very high (low) values of γ_{th} ($\eta_{\max}$) would inevitably lead to a performance degradation in terms of SNR NRMSE, due to the dominant contribution of the reflected signals to the overall received power. On the other hand, reflected rays of order higher than the second have a negligible impact in terms of SNR NRMSE (the gap between the $\eta_{\max} = 2$ and $\eta_{\max} = 3$ configurations, with $\gamma_{\text{th}} = -40$ dB, is just 0.001) in the face of a speedup improvement of around 100%. In this scenario, further reducing γ_{th} would result in a considerable increase of the system complexity while leading to negligible accuracy gain, and the optimal approach would be to select $\eta_{\max} = 2$ with $\gamma_{\text{th}} = -40$ dB.

Finally, we highlight that, while limiting the number of MPCs reduces the ray tracer complexity, it may preclude the implementation of beamforming techniques that exploit the sparsity property of the channel to realize simultaneous beams in independent angular direction (e.g., MIMO techniques exploiting spatial multiplexing). Additionally, while simplifications might have minimal implication on low-layer performance metrics, e.g., the SNR, their effect on higher-layer metrics, e.g., end-to-end throughput and latency, is still unknown and deserves further investigation.

V. Conclusions and Future Work

In this paper, we presented possible simplifications to reduce the complexity of channel modeling through ray tracing. Notably, after an introduction on how a RT works, and on which are the sources of computational complexity for the RT, we discussed two strategies which aim at avoiding computations for MPCs which do not contribute significantly to the overall received power. The first limits the maximum reflection order, while the second removes MPCs with a path gain which is much smaller than that of the strongest ray. We then evaluated the impact of these simplifications on the SNR, in three different scenarios, and on the run time of the RT and of a network simulator. We highlighted that, for each scenario, there exists an optimal working point which minimizes the accuracy loss with respect to the baseline, but reduces the channel generation and modeling time by up to 4 times.

As future works, we will consider a more complex RT, which also includes diffuse components, according to a quasi-deterministic model [20]. Moreover, we will study the impact of the simplifications on the higher layers of the protocol stack, by using the ns-3 802.11ad module [21] (which already integrates the RT) and by extending the ns-3 mmWave module [12] to use RT traces.

References

[1] 3GPP, “NR and NG-RAN Overall Description - Rel. 15,” TS 38.300, 2018.

[2] 3GPP, “New WID on extending current NR operation to 71 GHz,” Qualcomm, RP-193229 - 3GPP TSG RAN Meeting 86, December 2019.

[3] IEEE, “IEEE Standard for Information technology–Telecommunications and information exchange between systems–Local and metropolitan area networks–Specific requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band,” *IEEE Std 802.11ad-2012 (Amendment to IEEE Std 802.11-2012, as amended by IEEE Std 802.11ae-2012 and IEEE Std 802.11aa-2012)*, pp. 1–628, Dec 2012.

[4] ITU-R, “IMT Vision - Framework and overall objectives of the future development of IMT for 2020 and beyond,” Recommendation ITU-R M.2083, September 2015.

[5] S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges,” *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[6] S. Deng, G. R. MacCartney, and T. S. Rappaport, “Indoor and Outdoor 5G Diffraction Measurements and Models at 10, 20, and 26 GHz,” in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2016.

[7] M. Polese, F. Restuccia, A. Gosain, and J. e. a. Jornet, “MillimeTera: Toward A Large-Scale Open-Source MmWave and Terahertz Experimental Testbed,” in *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, ser. mmNets’19. Los Cabos, Mexico: ACM, 2019, p. 27–32.

[8] S. K. Saha, Y. Ghasempour, M. K. Haider, T. Siddiqui, P. D. Melo, N. Somanchi, L. Zakrajsek, A. Singh, R. Shyamsunder, O. Torres *et al.*, “X60: A programmable testbed for wideband 60 GHz WLANs with phased arrays,” *Computer Communications*, vol. 133, pp. 77–88, Jan. 2019.

[9] T. Bai and R. W. Heath, “Coverage and rate analysis for millimeter-wave cellular networks,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, February 2015.

[10] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, “Modeling and analyzing millimeter wave cellular systems,” *IEEE Transactions on Communications*, vol. 65, no. 1, pp. 403–430, Jan 2017.

[11] 3GPP, “Study on channel model for frequencies from 0.5 to 100 GHz,” 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.901, Jun 2018, version 15.0.0.

[12] M. Mezzavilla, M. Zhang, M. Polese, R. Ford, S. Dutta, S. Rangan, and M. Zorzi, “End-to-End Simulation of 5G mmWave Networks,” *IEEE Commun. Surveys Tuts*, vol. 20, no. 3, pp. 2237–2263, Third Quarter 2018.

[13] V. Degli-Esposti, F. Fuschini, E. M. Vitucci, M. Barbiroli, M. Zoli, L. Tian, X. Yin, D. A. Dupleich, R. Müller, C. Schneider, and R. S. Thomä, “Ray-Tracing-Based mm-Wave Beamforming Assessment,” *IEEE Access*, vol. 2, pp. 1314–1325, 2014.

[14] S. G. Larew, T. A. Thomas, M. Cudak, and A. Ghosh, “Air interface design and ray tracing study for 5G millimeter wave communications,” in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2013, pp. 117–122.

[15] A. Maltsev, A. Pudeyev, A. Lomayev, and I. Bolotin, “Channel modeling in the next generation mmWave Wi-Fi: IEEE 802.11ay standard,” in *22th European Wireless Conference*, May 2016.

[16] C. Lai, R. Sun, C. Gentile, P. B. Papazian, J. Wang, and J. Senic, “Methodology for Multipath-Component Tracking in Millimeter-Wave Channel Modeling,” *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 3, pp. 1826–1836, March 2019.

[17] P. Testolina, M. Lecci, M. Polese, M. Giordani, and M. Zorzi, “Scalable and Accurate Modeling of the Millimeter Wave Channel,” in *International Conference on Computing, Networking and Communications (ICNC)*, 2019.

[18] M. Polese and M. Zorzi, “Impact of Channel Models on the End-to-End Performance of mmWave Cellular Networks,” in *IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, June 2018.

[19] Z. Yun and M. F. Iskander, “Ray Tracing for Radio Propagation Modeling: Principles and Applications,” *IEEE Access*, vol. 3, pp. 1089–1100, 2015.

[20] C. Gentile, P. B. Papazian, R. Sun, J. Senic, and J. Wang, “Quasi-Deterministic Channel Model Parameters for a Data Center at 60 GHz,” *IEEE Antennas and Wireless Propagation Letters*, vol. 17, no. 5, pp. 808–812, May 2018.

[21] H. Assasa, J. Widmer, T. Ropitault, and N. Golmie, “Enhancing the Ns-3 IEEE 802.11ad Model Fidelity: Beam Codebooks, Multi-Antenna Beamforming Training, and Quasi-Deterministic MmWave Channel,” in *Proceedings of the 2019 Workshop on Ns-3*, ser. WNS3 2019. New York, NY, USA: Association for Computing Machinery, 2019, p. 33–40. [Online]. Available: https://doi.org/10.1145/3321349.3321354

Measurement of the magnet system for the QEMMS

R. Marangoni[1, 3], D. Haddad[1], F. Seifert[1, 2], L. Chao[1], D. Newell[1], and S. Schlamminger[1]
[1]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]University of Maryland, Joint Quantum Institute, College Park, MD 20742, USA
[3]E-mail: rafael.marangoni@nist.gov

***Abstract*—The magnet system for the Quantum Electro-Mechanical Metrology Suite (QEMMS) Kibble balance is being manufactured. The QEMMS Kibble balance will be used to measure masses nominally around 100 g and the magnet system will be able to generate forces up to 1 N. Some aspects of the magnet system and the measurement procedure used to determine the profile of the magnetic field are described here.**

***Index Terms*—QEMMS, Kibble balance, mass measurement, magnet system, gradiometer coil.**

I. Introduction

The Kibble balance is an instrument used to realize the unit of mass. In this instrument, the gravitational force generated by a mass prototype is compensated by the electromagnetic force of a current-carrying conductor in a magnetic field [1]. It is fundamental to measure the electromagnetic force with high accuracy in order to determine the mass of the prototype. By measuring the flux gradient Bl in the coil, it is possible to determine the relation between electromagnetic force and current. The flux gradient is equal to the product of the magnetic flux density in the coil region and the coil length. It is an advantage to have a flat Bl profile for determining the flux gradient in the coil region with high precision. In the following, a brief description of the magnet system, the electromagnetic force, and the measurement strategy used are provided.

II. Magnet system

The magnet system is shown in figure 1. It is composed of two permanent magnets and one iron yoke made of soft steel. The design of the magnet system is described in detail in [2]. The permanent magnets are comprised of TC-16 ($Sm_2Co_{17}Gd$) arc segments. This material exhibits a temperature coefficient for the remanent magnetization of $-0.001\,\%/\mathrm{K}$. The magnet system is responsible for generating the magnetic flux through the coil, which is necessary to produce the electromagnetic force. A stable remanent magnetization for the permanent magnets is necessary in order to perform the mass measurement with high precision.

III. Electromagnetic force

The original equations for the Kibble balance principle are shown in [1]. Consider a coil with inductance L which encloses a magnetic flux Φ_R. When a current I flows, the total magnetic flux through the coil is $\Phi = \Phi_\mathrm{R} + LI$ and the energy is $W = -I\Phi_\mathrm{R} - I^2L/2$. If Φ_R and L are a function of displacement z of the coil, then a component of force exists:

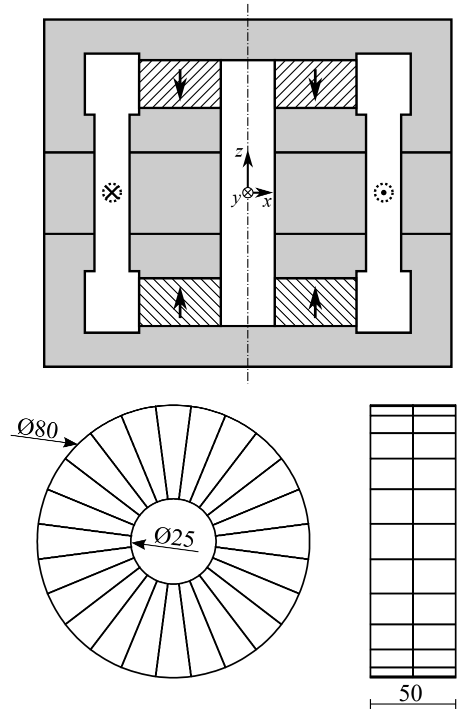


Fig. 1. Upper part: Magnet system with the coil. Axes x and y represent the horizontal directions and z axis represents the vertical direction. Lower part: Permanent magnet ring composed by TC-16 ($Sm_2Co_{17}Gd$) arc segments. The numbers are dimensions in millimeter.

$$F_\mathrm{z} = -\frac{\partial W}{\partial z} = I\frac{\partial \Phi_\mathrm{R}}{\partial z} + \frac{I^2}{2}\frac{\partial L}{\partial z} \tag{1}$$

If the coil moves with a velocity v in the same direction, an induced voltage U is obtained:

$$\frac{U}{v} = -\frac{\partial \Phi}{\partial z} = -\frac{\partial \Phi_\mathrm{R}}{\partial z} - I\frac{\partial L}{\partial z} \tag{2}$$

In the Kibble balance experiments, equation (1) is known as the force mode and equation (2) is known as the velocity mode. For the NIST Kibble balance, there is no current during the velocity mode (equation 2). In order to minimize measurement deviations caused by $\partial L/\partial z$ during the force mode, a symmetric weighing with mass-on/mass-off strategy is performed [2].

Marangoni, Rafael; El Haddad, Darine; Seifert, Frank; Chao, Leon; Newell, David B.; Schlamminger, Stephan. "Measurement of the magnet system for the QEMMS." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

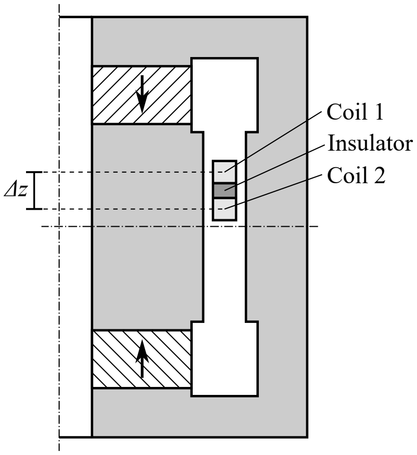


Fig. 2. Gradiometer coils used for measuring the magnetic field profile. Two coils separated by an insulator are used for the measurement.

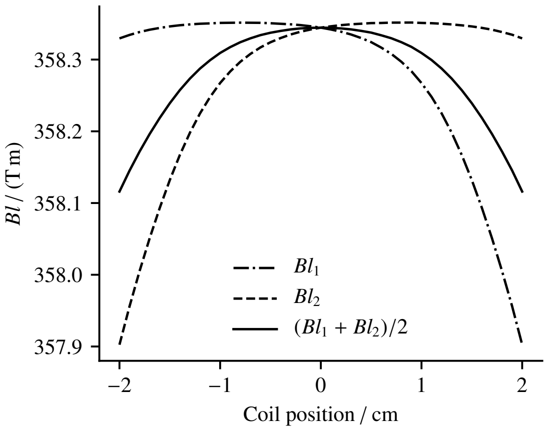


Fig. 3. Profile of the product Bl for the coils 1 and 2 and the mean value between both profiles. This is a result of finite element analysis.

The magnetic flux Φ_{R} is generated by a magnet system with permanent magnets. It is given as a function of coil position and angle x, y, z, θ_{x} and θ_y:

$$\Phi_{\mathrm{R}} = Bl\left(z + \frac{\theta_{\mathrm{y}}x + \theta_{\mathrm{x}}y}{2}\right) \tag{3}$$

where B is the radial flux density in the region where the coil resides and l is the total length of the coil. The coordinates x, y and z are given by a reference frame fixed to the center of the magnet system (figure 1). In a yoke based magnetic system, the coil inductance can be written as

$$L = L_0 - k_{\mathrm{z}}z^2 + k_{\mathrm{x}}x^2 + k_{\mathrm{y}}y^2 - k_{\theta\mathrm{x}}\theta_{\mathrm{x}}^2 - k_{\theta\mathrm{y}}\theta_{\mathrm{y}}^2 \tag{4}$$

where the quantities k_{z}, k_{y}, k_{x}, $k_{\theta\mathrm{y}}$ and $k_{\theta\mathrm{x}}$ are positive real numbers. The gradient of the flux and inductance in the vertical direction is related to the electromagnetic force component used to measure the mass. Ideally the weighing is performed in a position such that the gradient of the flux and inductance in the horizontal and angular directions can be neglected.

As previously mentioned, a flat Bl profile, or a region along z where $\mathrm{d}B/\mathrm{d}z$ approaches 0, is advantageous and can be achieved by implementing shimming techniques found in [3].

IV. Measurement strategy

The gradiometer coils shown in Fig. 2 are used to measure the magnetic flux profile. Two coils separated by a distance Δz of $15.9\,\mathrm{mm}$ are used. This distance is given between the center of both coils. When these coils move in the magnet system with a velocity v, the induced voltages U_1 and U_2 are obtained:

$$U_1 = -v\frac{\partial\Phi}{\partial z}\bigg|_{z+\Delta z/2} \tag{5}$$

$$U_2 = -v\frac{\partial\Phi}{\partial z}\bigg|_{z-\Delta z/2} \tag{6}$$

As described in [3], the profile of the magnetic field can be obtained by performing the following calculation:

$$\Delta Bl(z) = \frac{1}{\bar{v}\Delta z}\int_{z_0}^{z} U_1(\zeta) - U_2(\zeta)\mathrm{d}\zeta \tag{7}$$

The advantage of using this method to measure the Bl profile is the fact that it is not necessary to measure the velocity of the coil v with high precision. In this way, it is possible to use a simple setup to measure the profile.

Figure 3 shows the Bl profiles for the coils 1 and 2, indicated as Bl_1 and Bl_2 respectively. This is a simulation result obtained with finite element analysis. The mean value $(Bl_1 + Bl_2)/2$ between both profiles is shown in the same figure. As described in [2], a value of about $700\,\mathrm{T\,m}$ will be used for the magnet system and coil of the QEMMS. The gradiometer coils were designed in a way that it is possible to connect both coils in series and obtain the desired value for Bl. By doing so, it is possible to use the same coil for the profile measurement and balance operation.

V. Conclusion

The Quantum Electro-Mechanical Metrology Suite will be used to measure masses nominally around $100\,\mathrm{g}$ aimed at achieving absolute uncertainties around $2\,\mu\mathrm{g}$. The magnet system has been designed and is currently being manufactured. The gradiometer coil described in this paper will be used for measuring the profile of the magnetic field and ultimately for balance operation.

References

[1] B. P. Kibble, “A measurement of the gyromagnetic ratio of the proton by the strong field method,” in *Atomic Masses and Fundamental Constants 5 ed*, 1976, pp. 545–51.

[2] R. Marangoni, D. Haddad, F. Seifert, L. S. Chao, D. B. Newell, and S. Schlamminger, “Magnet system for the Quantum Electro-Mechanical Metrology Suite,” *IEEE Trans. Instrum. Meas.*, 2019, accepted for publication (DOI: 10.1109/TIM.2019.2959852).

[3] F. Seifert, A. Panna, S. Li, B. Han, L. Chao, A. Cao, D. Haddad, H. Choi, L. Haley, and S. Schlamminger, “Construction, Measurement, Shimming, and Performance of the NIST-4 Magnet System,” *IEEE Trans. Instrum. Meas.*, vol. 63, no. 12, pp. 3027–3038, Dec 2014.

Design of the Kibble balance for the QEMMS

R. Marangoni[1, 3], D. Haddad[1], F. Seifert[1, 2], L. Chao[1], J. Pratt[1], D. Newell[1], and S. Schlamminger[1]
[1]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]University of Maryland, Joint Quantum Institute, College Park, MD 20742, USA
[3]E-mail: rafael.marangoni@nist.gov

***Abstract*—The design status of the Kibble balance for the Quantum Electro-Mechanical Metrology Suite is provided. The balance is being developed with the objective to obtain a simple and robust design while performing high precision measurements. Aspects related to the vacuum vessel and balance mechanism are described. This includes some design information and simulation results.**

***Index Terms*—QEMMS, Kibble balance, mass measurement, vacuum chamber, balance mechanism**

I. Introduction

The Kibble balance is used to realize the kilogram unit since last revision of the International System of Units (SI) [1]. According to the SI, the unit of mass can be realized based on the Planck constant, the speed of light in vacuum and the hyperfine transition frequency of the caesium-133 atom. The Quantum Electro-Mechanical Metrology Suite (QEMMS) is composed by a Kibble balance, a Josephson voltage standard and a Quantum Hall resistance standard. With this system, it is possible to measure mass, length, time, voltage, electrical resistance and current traceable to the SI. The Kibble balance is being developed for measuring masses of 100 g with an uncertainty of 2 µg. The magnet system of the balance has already been designed is currently being manufactured [2]. Further aspects include the balance mechanism and the vacuum chamber, which are discussed in the following.

II. General aspects of the design

The Kibble balance of the QEMMS is being designed to be robust, precise and simpler than the NIST-4 [3]. It will be smaller and will have less operational requirements. Figure 1 shows the general design of the balance with the mechanism, permanent magnet, vacuum chamber, vacuum pump, test mass, counterweight and magnet system for the counterweight side. Some ideas of mechanism and pivots are described in [4]. There are different choices for the mechanism and the pivot, most of them using flexure strips or knife edges.

The balance mechanism is responsible for moving the coil during the velocity mode and acting as a weighing mechanism in the force mode. For the force mode, it is desired to have a stiffness as small as possible. The position of the coil in the vertical direction is measured by using a single laser interferometer. A corner cube located in the center of the coil is used for that. The laser beam (not shown in the figure) is launched outside the balance under the corner cube.

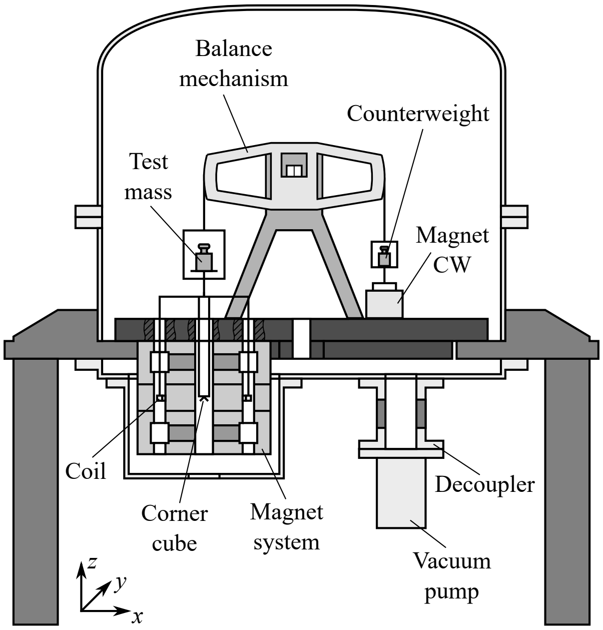


Fig. 1. Design for the QEMMS Kibble balance including the balance mechanism, vacuum chamber and magnet system.

The balance is operated in vacuum in order to avoid measurement deviations caused by the buoyancy force of air in the mass prototype, and also deviations caused by the refractive index of air during the length measurement. For this reason, a vacuum chamber is necessary.

III. Vacuum chamber

A drawing of the vacuum chamber is shown in figure 2. It is basically composed by four parts: dome, base, vessel bottom and bucket. The vessel is designed to operate under pressures down to 1×10^{-4} Pa. The components are made from 304 stainless steel and the inside of the chamber is electropolished. The bucket has to be removed in order to obtain access to the magnet system of the balance. As described in [2], the magnet system was designed with the objective to be split in situ. By removing the dome, it is possible to obtain access to the balance mechanism. This will be done by using a motorized hoist. As shown in figure 1, the vacuum pump is located under the balance and an electrical decoupler is used to mount the pump to the balance.

Marangoni, Rafael; El Haddad, Darine; Seifert, Frank; Chao, Leon; Pratt, Jon R.; Newell, David B.; Schlamminger, Stephan. "Design of the Kibble balance for the QEMMS." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

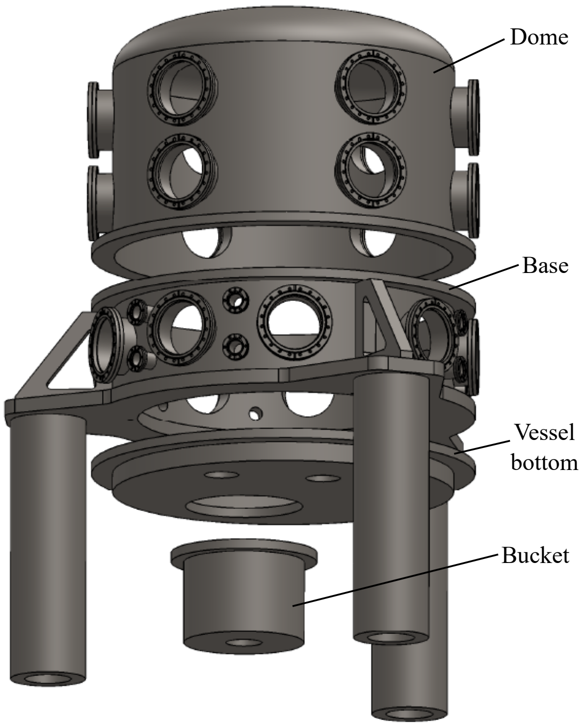


Fig. 2. Vacuum chamber for the QEMMS Kibble balance. The balance is composed by four parts: dome, base, vessel bottom and bucket.

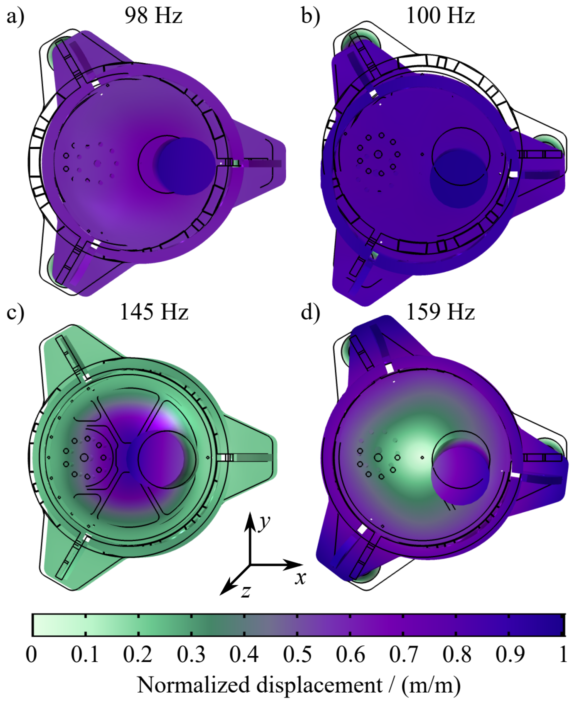


Fig. 3. Vibration modes of the vacuum chamber and respective eigenfrequencies: a) vibration mode along x; b) vibration mode along y; c) drum vibration mode; d) torsional vibration mode about z. Color scale shows the normalized displacement.

Figure 3 shows a result of finite element analysis used to determine the vibration modes and the frequencies of vibration of the vacuum chamber. The dome is not shown in this figure. Since the balance mechanism and its weight are not exactly known yet, an equivalent mass with an estimated weight value was added to the simulation model. The lowest frequencies of vibration are given by the vibration modes in the x and y directions. The natural frequencies are 98 Hz and 100 Hz respectively. The drum vibration mode, which is mostly present in the base plate of the balance, has a natural frequency of 145 Hz. The torsional vibration mode about the z direction has a frequency of 159 Hz. These frequencies are much higher when compared to the frequencies of the balance mechanism. By doing so, the influences of vibrations of the vacuum chamber in the balance mechanism are reduced.

A preliminary design of the vacuum chamber that fulfills the operational requirements of the balance was performed. The size of the vacuum chamber for the QEMMS Kibble balance is at least 7 times smaller in volume than the NIST-4 vacuum chamber. The design itself is simpler and avoids the use of bellows. With the new design, it is possible to access the magnet system without disassembling the balance. The vacuum pump can be assembled under the balance and, by doing so, a more compact design is obtained.

IV. Conclusion

The Kibble balance of the Quantum Electro-Mechanical Metrology Suite (QEMMS) will be used to realize the kilogram unit in a range under 100 g with high precision. The magnet system of the balance was designed and is being manufactured. The vacuum chamber and the balance mechanism are currently being designed. Some simulation results used to determine the vibration modes of the vacuum chamber and the natural frequencies are shown.

References

[1] I. A. Robinson and S. Schlamminger, “The watt or Kibble balance: a technique for implementing the new SI definition of the unit of mass,” *Metrologia*, vol. 53, no. 5, pp. A46–A74, sep 2016.

[2] R. Marangoni, D. Haddad, F. Seifert, L. S. Chao, D. B. Newell, and S. Schlamminger, “Magnet system for the Quantum Electro-Mechanical Metrology Suite,” *IEEE Trans. Instrum. Meas.*, 2019, accepted for publication (DOI: 10.1109/TIM.2019.2959852).

[3] D. Haddad, F. Seifert, L. S. Chao, S. Li, D. B. Newell, J. R. Pratt, C. Williams, and S. Schlamminger, “Invited article: A precise instrument to determine the planck constant, and the future kilogram,” *Review of Scientific Instruments*, vol. 87, no. 6, p. 061301, 2016.

[4] R. Marangoni, D. Haddad, F. Seifert, L. S. Chao, D. B. Newell, and S. Schlamminger, “Design of the QEMMS Kibble balance,” in *Proceedings of the 34th ASPE Annual Meeting*, vol. 71, 2019, pp. 166–70.

Marangoni, Rafael; El Haddad, Darine; Seifert, Frank; Chao, Leon; Pratt, Jon R.; Newell, David B.; Schlamminger, Stephan. "Design of the Kibble balance for the QEMMS." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

Calibration of an AC Voltage Source Using a Josephson Arbitrary Waveform Synthesizer at 4 V

Nathan E. Flowers-Jacobs[1], Alain Rüfenacht[1],
Anna E. Fox[1], Paul D. Dresselhaus[1], and Samuel P. Benz[1]
[1]National Institute of Standards and Technology, Boulder, CO 80305, USA
nathan.flowers-jacobs@nist.gov

***Abstract*—This paper describes a method for calibrating an ac source using a Josephson Arbitrary Waveform Synthesizer (JAWS) by summing the sources in series and tuning the magnitude and phase of the JAWS to null the combined output voltage. The method requires an ac source that can generate a signal that is phase-synchronous with the JAWS. As a demonstration of this method, we measure the output of a calibrator at an rms output of 4 V and a frequency of 1 kHz.**

***Index Terms*—Digital-analog conversion, Josephson junction arrays, measurement standards, signal synthesis, superconducting integrated circuits, voltage measurement.**

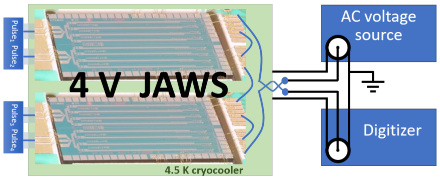


Fig. 1. Diagram of the JAWS source (low-frequency biases are not shown), an ac voltage source under test, and a digitizer used to determine the residual difference between the voltages generated by the JAWS and the source.

I. Introduction

AC voltage metrology is currently based on thermal transfer standards. These standards are broadband rms detectors that typically achieve better than 10 μV/V accuracy and 1 μV/V repeatability in determining the voltage generated by an ac source at voltages less than 10 V and frequencies less than 100 kHz [1]. However, rms detectors are inherently unable to provide information about the spectral purity and phase stability of a source; thus, the measured value incorporates the rms contribution of all input signals spread over the greater-than 10 MHz bandwidth of the detector.

We can improve on this approach based on thermal rms-detectors by directly comparing ac voltage source signals to quantum-based ac voltages generated by a Josephson Arbitrary Waveform Synthesizer (JAWS). A JAWS-based approach results in quantum-based calibrations that are directly linked to the revised SI, and offers a different paradigm of source-based, instead of detector-based, ac metrology. It could also lead to shorter calibration times and improved accuracy.

The basic measurement setup is effectively an ac bridge measurement, shown in Fig. 1. The JAWS source is in series with the output of the ac voltage source under test, and a high-impedance digitizer is used to measure the summed voltage. Since the JAWS is a floating voltage source, it is placed on the "high" side of the ac voltage source and digitizer. The "low" side of the ac voltage source and digitizer are grounded.

Using this setup, the magnitude and phase of the quantum-based JAWS source is tuned to null the voltage measured at the digitizer. Use of a null measurement reduces the importance of the digitizer calibration, and makes it significantly easier to measure small changes in the relative phase and magnitude over time. It is also crucial to confirm that the JAWS source is operating correctly during the measurement. There are two steps to this test. First, it is important to measure the full quantum-locked range (QLR) of the JAWS at the required output voltage magnitude and phase, that is, to confirm that the output of the JAWS source is independent of the JAWS bias parameters over a reasonable range of bias parameters. Second, mini-QLRs should be taken during the ac voltage comparison by making changes to the JAWS bias parameters, which are small enough that the changes are not expected to affect the JAWS output but large enough to observe problems [2].

II. AC Source Requirements

This null measurement approach requires that the ac source be able to operate in a phase-synchronous manner with the JAWS source. There are a number of different ways this requirement can be achieved. One approach is to have the JAWS and ac source share a fast clock signal, typically at 10 MHz, which ensures that the two instruments have synchronous waveforms. Then the two signals can be phase-synchronized by having one system provide a trigger signal at the beginning of a waveform period. Another approach is to force the ac voltage source to phase-lock to a signal provided by the JAWS system at the frequency of interest.

The key feature of both of these approaches, and the requirement for other phase-locking schemes, is that the magnitude and phase of the JAWS output voltage can be changed relative to the ac voltage source in a stable, repeatable manner.

III. Preliminary Results and Discussion

As a demonstration of this technique, we use a JAWS to measure a Fluke 5700A calibrator[1] generating an rms voltage

[1] Commercial instruments are identified in this paper in order to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the equipment identified is necessarily the best available for the purpose.

Flowers-Jacobs, Nathan; Rufenacht, Alain; Fox, Anna; Dresselhaus, Paul; Benz, Samuel P. "Calibration of an AC Voltage Source Using a Josephson Arbitrary Waveform Synthesizer at 4 V." Paper presented at Conference on Precision Electromagnetic Measurements, Denver, CO, US. August 24, 2020 - August 28, 2020.

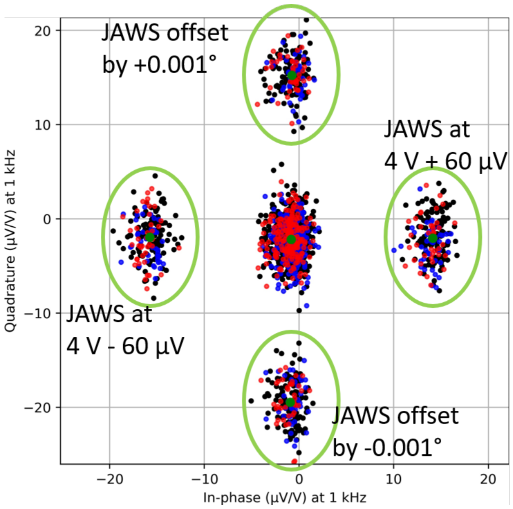


Fig. 2. Residual difference between the 1 kHz 4 V waveforms generated by the JAWS and a calibrator over 13 hours, plotted in the complex plane. The difference after optimizing the JAWS output is in the center (not circled). The JAWS output is also offset by ±0.001°or ±60 μV (labeled green circles). We also take mini-QLRs during the measurement: a JAWS DC bias current of 0 mA (black), +0.5 mA (blue) or -0.5 mA (red) is applied to test the JAWS operation.

of 4 V at 1 kHz. To phase-lock the two systems, the JAWS source provides a stable, optically isolated, 1 kHz square wave to the ”PHASE LOCK IN” port of the calibrator. The phase of the JAWS source is then tuned relative to this square wave, along with the magnitude of the JAWS source, to minimize the voltage measured at 1 kHz on a Zurich Instruments MFLI digitizer. After optimizing the JAWS source, the dc offset current QLR of the JAWS is greater than 1.4 mA with the calibrator generating 4 V and the JAWS generating 4 V in opposition at 1 kHz.

In Fig. 2, we show the results of this measurement over 13 hours. Each data point is the result of analyzing 0.5 s of data. The data are plotted in the complex plane, with small differences in phase along the quadrature/vertical axis and small differences in amplitude along the in-phase/horizontal axis. We also changed the JAWS output by ±0.001°and ±60 μV to characterize the digitizer. Finally, we performed mini-QLRs by applying a dc bias offset of ±0.5 mA to the JAWS arrays during some of the measurements. Since we did not observe any resolvable effects or trends due to these bias variations, we conclude that JAWS produced accurate voltages during the entire measurement. Using all of these data, we note that the calibrator produces a 4 V signal that is, on average, +0.8 μV/V or 3 μV larger than 4 V.

This technique can also be used to look at the short-term phase stability of ac voltage sources. In Fig. 3 we show the results of analyzing in finer detail the data that were used to create a single point in Fig. 2. In Fig. 3, we separately plot the in-phase and quadrature data as a function of time; each data point is taken from four periods of the 1 kHz waveform. We detect short-term changes in the relative phase of the systems which are both significantly larger than the amplitude changes and clearly resolvable over the system noise; similar phase jitter has been seen before [3].

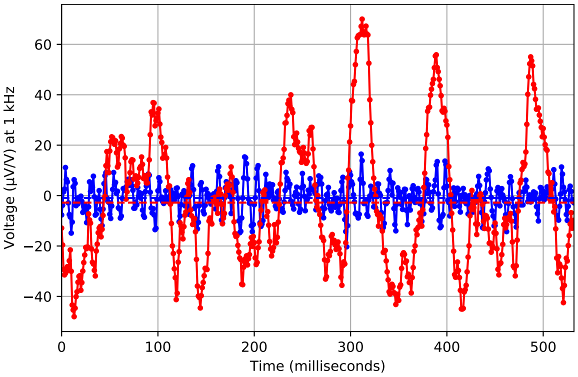


Fig. 3. Residual difference between the 1 kHz 4 V waveforms generated by the JAWS and calibrator plotted versus time. The short-term fluctuations in the quadrature component of the data (red, proportional to phase) are significantly larger than those in the in-phase component (blue, proportional to amplitude).

These results demonstrate the important aspects of this measurement method. At CPEM 2020, we will provide more results and a more detailed uncertainty analysis for evaluating phase and amplitude stability of ac sources. We will also provide results of measurements taken over a wide range of amplitudes and frequencies.

A similar type of analysis can also be used to characterize distortion in ac signals by either directly using the higher frequency information gathered by the digitizer or by distorting the JAWS waveform to cancel the higher frequency content generated by the ac source. The first approach may require calibration of the digitizer, but digitizer non-linearity is less important because the fundamental tone is the predominate signal, and it has already been nulled by the JAWS signal in opposition. The second approach removes any dependence on the digitizer calibration, but requires that the distortion produced by the ac source be stable.

IV. Conclusion

We demonstrated that a quantum-based JAWS voltage source can be used to calibrate other ac voltage sources using a null measurement method where the magnitude and phase of the JAWS is tuned to cancel the other ac voltage source. In the future, this approach can also be used to investigate the short-term stability and distortion of ac voltage sources.

References

[1] J. R. Kinard, J. R. Hastings, T. E. Lipe, and C. B. Childers, “AC-DC Difference Calibrations,” *NIST Special Publication*, vol. 250-27, 1989.

[2] A. Rüfenacht, N. E. Flowers-Jacobs, and S. P. Benz, “Impact of the latest generation of Josephson voltage standards in ac and dc electric metrology,” *Metrologia*, vol. 55, no. 5, pp. S152–S173, Oct 2018.

[3] A. Rufenacht, C. J. Burroughs, P. D. Dresselhaus, and S. P. Benz, “Differential sampling measurement of a 7 V rms sine wave with a programmable Josephson voltage standard,” *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 6, pp. 1587–1593, June 2013.

**Proceedings 2020 ASME International Manufacturing Science and Engineering Conference
MSEC2020
June 22-26, University of Cincinnati, USA**

MSEC2020-8281

ADVANCED SENSING DEVELOPMENT TO SUPPORT ACCURACY ASSESSMENT FOR INDUSTRIAL ROBOT SYSTEMS

Guixiu Qiao
National Institute of Standards and Technology
Gaithersburg, Maryland, USA

Jonathan Garner
University of Maryland
Collage Park, Maryland, USA

KEY WORDS

Advanced Sensing, Condition Monitoring, Diagnostics, Prognostics Maintenance, Health Management, Industrial Robot Systems, Performance Degradation

ABSTRACT

Manufacturers currently struggle with the assessment of a machine/robots' accuracy degradation that limits the efficiency of machine/robots in high precision applications. Current best practice in industry is to inspect the final products or add redundancies (local calibration, etc.) during the process to determine the machine's accuracy and performance. These create complexities in the process and increase the maintenance costs of applications such as high precision robot operations (welding, robotic drilling/riveting, and composite material layout), in-process metrology, and machines in mobile applications. A higher speed, more precise control of position and orientation is required to remedy these complexities. A novel smart target was designed at the National Institute of Standards and Technology (NIST) to integrate with a vision system to acquire high-accuracy, real-time 6-D (six-dimensional x, y, and z position, roll, pitch, and yaw orientation) information. This paper presents the development of the smart target and the image processing algorithm to output 6-D information. A use case is presented using the smart target on universal robots (UR3 and UR5) to demonstrate the feasibility of using the smart target to perform the robot accuracy assessment.

INTRODUCTION

Machines and robots are key automation instruments that are widely used in manufacturing, material handling, construction, medicine, and aerospace [1]. In recent years, robots have become more accurate due to improvements in motion control, actuators, and other technologies [2]. These improvements enable the broader use of robots in many new applications. Machines and robots' accuracy assessment is crucial to these applications. Unexpected disturbances due to accuracy degradation may lead to a degradation of the system performance, causing losses in productivity and business opportunities [3]. With current industry practice, it is difficult to detect the accuracy degradation because the machine or robot is continuously running and appears to be making parts that wouldn't not meet the quality requirements, including accuracy requirement.

As shown in Fig. 1, the typical robot errors contain static errors and dynamic errors. Static errors of a 6-axis robot include geometric error (e.g., linkage length, tools, and object in workspace), elasticities (e.g., base, runout, and gears), and temperature change created errors (quasi-static errors); dynamic errors include the trajectory following errors, gear cyclic errors, and axis dynamic limits [2, 4]. Because these errors influence robots' absolute accuracy, traditional hard automation must depend on robot teaching, which is very time-consuming thus increasing the costs of manufacturing. To work around the accuracy problems, many extra sensors and redundancies are added. For example, flexible grippers or extra sensors are implemented to increase task tolerance. Local calibrations are developed to improve local accuracy for the success of precision operations. Sometimes external guidance, for example, a laser tracking system, is added to guide the robot's precision operations [5, 6]. These workaround methods significantly increase the complexity of the manufacturing system and the cost of system maintenance. Moreover, there are new emerging robot applications that require more precise robot operations, for example, high precision assembly, welding, robotic drilling/riveting, robot metrology, and composite material layout. There needs to be an innovative way to design new robot systems instead of continuously using 20-year-old methods. The robot accuracy needs to be measured, assessed, monitored, and improved to support the development of an optimized and simplified production line by enhancing the absolute accuracy. Moreover, economic factors also motivate facilities and factories to perform accuracy degradation assessment and monitoring the robot performance to detect faults and failures. The purpose is to improve maintenance techniques and operations, especially eliminating unexpected shutdowns.

Qiao, Guixiu; Garner, Jonathan. "Advanced sensing development to support accuracy assessment for industrial robot systems." Paper presented at ASME International Manufacturing Science and Engineering Conference, Cincinnati, OH, US. June 22, 2020 - June 26, 2020.

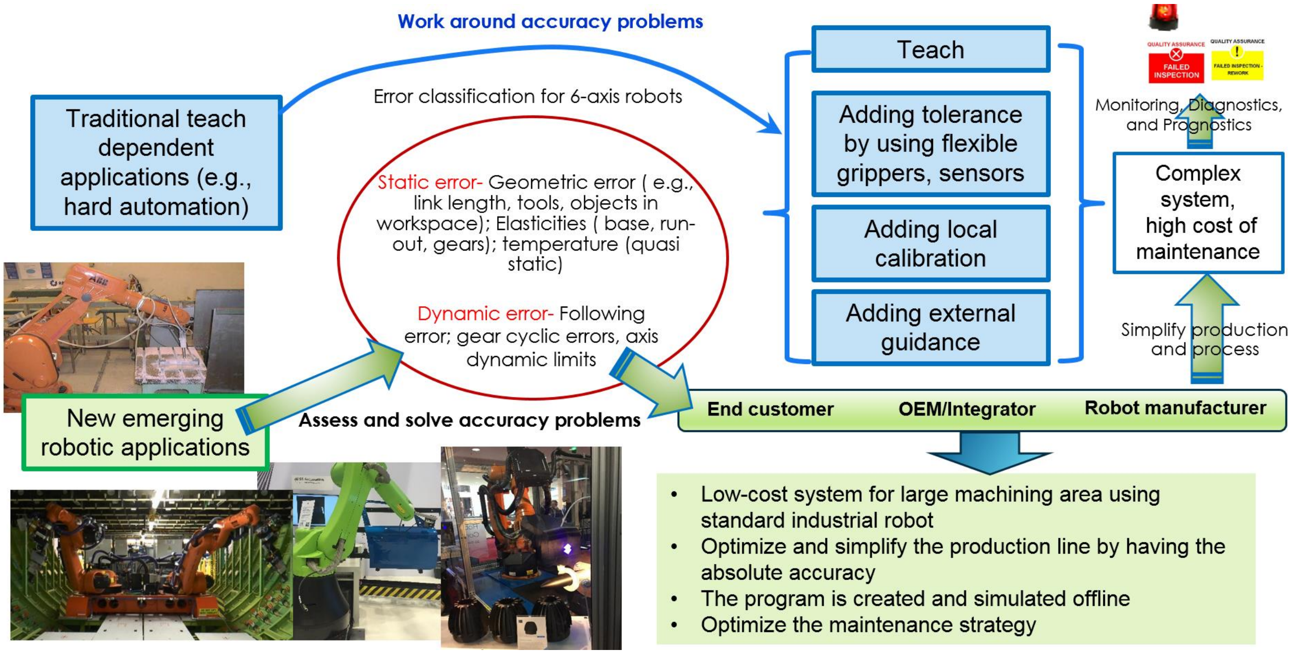

Work around accuracy problems
Traditional teach dependent applications (e.g., hard automation)
Error classification for 6-axis robots
Static error- Geometric error (e.g., link length, tools, objects in workspace); Elasticities (base, run-out, gears); temperature (quasi static)
Dynamic error- Following error; gear cyclic errors, axis dynamic limits
New emerging robotic applications
Assess and solve accuracy problems
Teach
Adding tolerance by using flexible grippers, sensors
Adding local calibration
Adding external guidance
QUALITY ASSURANCE
FAILED INSPECTION
QUALITY ASSURANCE
FAILED INSPECTION - REWORK
Monitoring, Diagnostics, and Prognostics
Complex system, high cost of maintenance
Simplify production and process
End customer
OEM/Integrator
Robot manufacturer
Low-cost system for large machining area using standard industrial robot
Optimize and simplify the production line by having the absolute accuracy
The program is created and simulated offline
Optimize the maintenance strategy

Figure 1. Robot accuracy degradation and influences

A robot's position (x, y, z) and orientation (pitch, yaw, roll) need to be measured to assess the robot's accuracy. The measured 6-D information can be used to calculate the deviations of robot position and orientation, allowing for accuracy degradation monitoring of a robot. The data can also be used as feedback for more accurate control, or as the input of an algorithm to perform calibration for robot performance improvement.

A novel smart target (patent pending) was developed at the National Institute of Standards and Technology (NIST) to integrate with a vision system to acquire high accuracy 6-D information of a moving object. The smart target is mounted on the object of interest, for example, the end effector/tool of a robot arm or the last link of a machine tool to measure/track the object's 6-D position and orientation. The smart target development is a part of the Prognostics and Health Management for reliable operations in Smart Manufacturing (PHM4SM) research at NIST. The PHM4SM project works on developing and deploying measurement science to promote the implementation, verification, and validation of advanced monitoring, diagnostic, and prognostic technologies to increase reliability and decrease downtime in smart manufacturing systems. This paper will present the research background, advanced sensing development, and a software tool to efficiently measure, monitor, diagnose, predict, and maintain the health of a robot. A use case was developed at NIST using Universal Robot UR3 and UR5 to demonstrate the feasibility of the smart target in accuracy assessment application.

RESEARCH BACKGROUND AND APPROACH

There are various measurement systems to acquire 3-D/6-D information [7, 8]. Some old methods including gauges, pose matching, and coordinate measurement machines are very slow. Trilateration with a theodolite, using cable potentiometer systems, or laser interferometers and other methods usually lack orientation information [9-13]. The laser tracker and vision-based system are gaining more attention in recent years.

Laser trackers are one type of high precision 3-D/6-D measurement system [18]. As an important part of the measurement system, measurement targets define what dimensional information can be captured by the system. If a 3-D target is used, 3-D information is captured. If a 6-D target is used, 6-D information is captured. Retro-reflective spheres are an example of a 3-D target for laser trackers. A laser tracker tracks the target to measure distance from the reflected laser beam. Encoders on the laser tracker provide two angular orientations of the tracker's two mechanical axes. By combining the distance and two encoders' angles, the center (x, y, z) of the retro-reflective target is measured [14, 16]. For 6-D information measurement, extra sensors are added to the existing 3-D target, for example, multiple light-emitting diodes (LEDs), thus making a already-expensive system more complex. The retro-reflective target needs to be held in contact with the object of interest. Also, laser tracker systems need to maintain line-of-sight between the laser tracker and the target. This means that the tracker will ultimately lose its view of the target when observing the target on a robot rotating to an angle.

Vision-based systems have the advantages of cost-effectiveness and broader application potentials with advanced image processing technology. The vision-based system includes camera array and structured light technology.

The camera array approach uses multiple cameras placed at different positions to capture multiple images of the same target [17, 19]. A dual-camera system is the simplest yet most popular camera array. Two cameras are separated by a distance, usually with a similar angle of view to benefit the disparity calculation. For each point in space, there is a measurable disparity between its positions in the two camera images. The depth of the point is then calculated using geometry. The main challenge of a camera array is how to find matching points in multiple images with good accuracy. Non-ideal point matching decreases the system's accuracy. Particularly, when multiple cameras are placed with a certain angle to enlarge the overlap of cameras' imaging for larger measurement capability, the non-ideal point matching problem gets more severe. Also, when the measured parts do not have enough features, for example, a smooth surface, the measurement accuracy decreases.

Structured light is an example of a special version of a camera array. Additional active projectors are added to solve the image matching problem in camera array. Projected patterns include fringes, random pattern laser dots, or other known patterns. These known patterns created a phase map to match the matching points in multiple images. A receiver detects the distortion of the reflected pattern to calculate a depth map based on geometry. The structured light system has better depth accuracy performance compared to the camera array system. However, the structured light system may be more expensive because of the costs of active projectors.

Both the camera array and structured light system are sensitive to environmental light. For the camera array, a bright environment works best. Structured light systems work best in a dark environment. When the brightness of the environment changes, images captured by a camera array may become noisy, and contrast becomes poor. This makes point matching extremely difficult resulting in inaccurate depth estimates. Moreover, structured light systems usually need to scan through a set of projected patterns. The measurement instrument and the measured part need to stay stationary during the measurement, which is not suitable for dynamic measurement.

Thus, although a variety of 6-D measurement systems and targets are available, these conventional systems do not have acceptable accuracy and dynamic features that are sufficiently accurate as required by some applications. As such, a new smart target was developed at NIST working with vision-based systems to overcome the challenges presented by complex industrial environments, enabling the measurement of dynamic poses.

ADVANCED SENSING DEVELOPMENT

The smart target system (patent pending) is a novel design to exceed the performance of existing vision-based measurement systems, especially with respect to accuracy and real-time processing potential. As shown in Fig. 2, the smart target consists of fixed-wavelength light pipes and two high-precision rotary gimbals. Three cylindrical light pipes, each a different color, are used to define line features that construct the 6-D information of a coordinate frame. The fixed-wavelength design makes the target stand out from an industrial background. At the same time, the target is not sensitive to environmental light. The red cross allows the vision system to detect the cross center as a coordinate origin. The gimbals are motorized to constantly rotate the red cross toward the measurement instrument for non-blocking dynamic measurement. It maximizes the target's line-of-sight to the vision system, thereby reducing measurement uncertainty. The blue and green pipes move with the object of interest and allow the camera system to determine orientation. This novel design enhances the matching of features across multiple images, especially in the presence of complex, industrial backgrounds. The smart target is mounted on the object of interest, for example, the end effector or tool of a robot arm, or the last link of a machine tool to measure and track the object's 6-D position and orientation. The smart target provides:

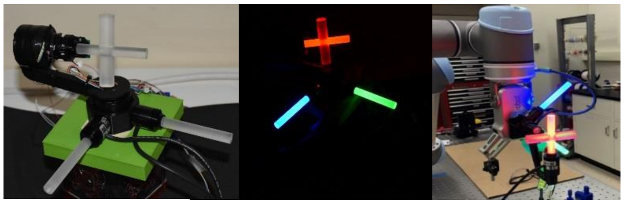


Figure 2. NIST designed smart target

1) High accuracy. Traditional targets have large uncertainty in measuring orientation. The most common traditional targets for vision systems are spheres. With infrared camera systems, the spheres are coated with reflective material or wrapped with reflective tapes. The center of the sphere is the feature to be measured. Multiple spheres are put together to define a coordinate frame. One sphere center may be used to define the origin of the coordinate frame. An axis is defined by two spheres centers. For this type of target, the measurement uncertainties of the sphere center are transferred one-to-one to the coordinate origin definition. Since only two points are used to define the axial direction, the angle measurement uncertainties are enlarged since a small distance error of the sphere center can create a large angular error. On the contrary, the axial direction of the smart target is defined using many points along the cylindrical target's centerline. Thus, the constructed center line is more accurate by fitting multiple points instead of only two points. For the same reason, the origin of the coordinate is created by intersecting two centerlines, which leads to an accuracy increase of 3 times compared to the traditional method of defining the origin using a sphere center. Moreover, extra features of the light pipe, including the fixed-wavelength color, the edge features of cylinders, etc., give more redundant information to improve the line detection accuracy.

2) Non-blocking measurement design to measure both static and dynamic robot tool center point (TCP) data. Traditional targets have the problem of bad pose (perpendicular to the

camera) when the target pose is not sensitive to camera measurement, or the target may block itself in some poses. The smart target has the red cross mounted on rotary axes. The red cross can constantly rotate toward the measurement system. The red cross center is defined as the coordinate system's origin. The rotation mechanism makes the smart target's origin good for measurement in different views without self-blocking.

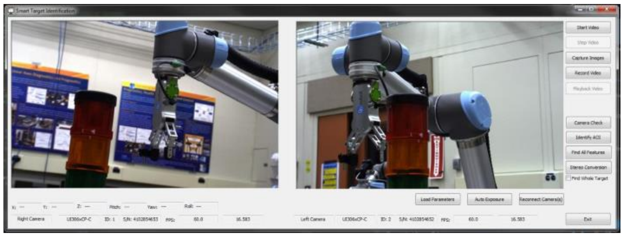
3) A unique definition of a coordinate frame. Traditional spherical targets do not have a unique definition of a frame. With the bad pose problem, spheres that define the origin may be blocked for measurement. Traditional sphere targets usually use best-fit transformation to find translation and rotation of two sets of center points. Best-fit usually uses the minimum least square errors for the conversing condition, thus not guaranteeing the consistent and minimum error for the origin. The 6-D smart target has a consistent and unique definition of a frame to avoid confusion when multiple coordinates exist in a system.

The 6-D smart target allows the continuous measurement of the 6-D information of a moving object with high accuracy. Applications of the smart target can be any general measurement system that requires high accuracy 6-D information of a moving object, for example, the robot and machine calibration, or multiple machines/tools/objects registrations, or adaptive objects location for unplanned adaptive control, or precisely tracking the pose of an object.

SOFTWARE TOOL DEVELOPMENT

A software tool is needed to process smart target-captured images, extract features, and output 6-D data. Software development was divided between creating a graphical user interface (GUI) to interface with the stereo camera system and designing and implementing an image processing algorithm to identify the smart target and determine its position using stereo images.

The GUI interfaces with the stereo camera system to allow for image processing is shown in Fig. 3. The software implements basic features such as image capture from each camera, video recording from one or both cameras, video playback, live video feeds, and smart target identification. An event handler responds to cameras connecting and disconnecting, as well as image transfer including transfer errors. The software receives images from the two cameras, displays them, and processes them as the user desires.



Smart Target Identification

X: — Y: — Z: — Pitch: — Yaw: — Roll: —

Right Camera UI306xCP-C ID: 1 S/N: 4102854653 FPS: 60.0 16.583

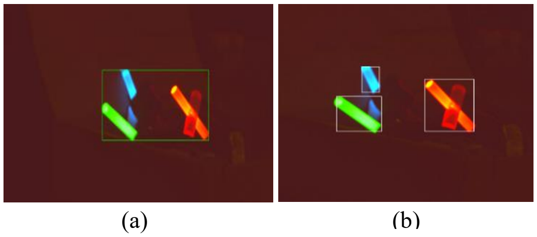
Left Camera UI306xCP-C ID: 2 S/N: 4102854652 FPS: 60.0 16.583

□ Find Whole Target

Figure 3. GUI to interface with the stereo camera system

While the GUI is receiving images from the cameras, the smart target can be tracked. An area of interest is drawn around the smart target in the live video feeds on the GUI to indicate the portion of the image to be used for further processing for location and orientation.

The image processing algorithm uses multiple layers of processing to identify the smart target and extract its location in 3D space. A software filter is applied to images from the two cameras to accentuate the red, blue, and green colors of the smart target. The filter accentuates both the horizontal and vertical cylinders of the red cross so the two can be distinguished. In each image, an area of interest (AOI) is found that encompasses the entire target as shown in Fig. 4 (a). The AOI reduces the size of the frame that must be further processed to reduce processing time for each frame.

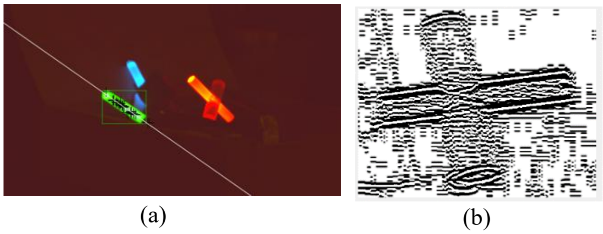


(a) (b)

Figure 4. Identify target from complex background

To find the AOI, a threshold filter is applied to identify the bright smart target. A morphological opening is also applied to reduce noise and to remove unwanted pixels that pass through the threshold filter. The detected AOI is drawn on top of live video feed on the GUI to display the part of the frame that will be processed further. To identify the three parts of the smart target and differentiate them based on color, a HSV filter is used. The three parts of the smart target can be identified (as shown in Fig. 4 (b)).

The next important part of the image processing is to identify the center lines of the green and blue cylinders and the center point of the red cross. The center of the red cross is found by finding the intersection of the lines running through the center of the red cross. A Laplacian of Gaussian (LOG) filter is applied to each colored part of the smart target. The filter marks the edges of the cylinder by a sharp jump from negative intensity to positive intensity. The zero crossing on each edge can be found and from there the centerline. Zero crossings on either side of the cylinder correspond with a point along the center of the cylinder. A line of best fit is found through the center points (Fig. 5 (a)). The center points, shown in black in Fig. 5 (a), come from the



(a) (b)

Figure 5. Feature detection of the smart target

edges of the cylinder. The line of best fit, shown in white in Fig. 5 (a), is found from the center points.

The red cross presents a challenge due to its more complex shape and unequal lighting. One cylinder of the red cross is evenly lit while the other is dim and uneven. This requires different processing techniques for the two parts. The evenly lit section of the cross is processed the same way as the blue and green cylinders using a LOG filter. The output of the Laplacian of Gaussian filter is shown in Fig 5 (b). The bright cylinder of the cross is the horizontal cylinder in Fig. 5 (b), and the dark cylinder is the vertical cylinder.

After the LOG filter, the two edges of the cylinder are found, then the center points, and the line of best fit of the centerline. The darker cylinder of the red cross is processed using a Canny filter to clearly identify the weaker edges of the vertical cylinder as shown in Fig. 6 (a). The Canny filter is not used in place of the LOG filter in other parts of the processing as it is prone to noise and weak edges found in the background. The LOG filter is more effective at singling out the edges of the bright cylinders. The results of the two processing methods for the cross are combined to find the center point of the red cross (Fig. 6 (b)). To determine the smart target position in 3D space, the location of the red cross center point, and the line of best through the green and blue cylinders are used from the two stereo images. The image pixel coordinates of the red cross and equations of the two lines are undistorted and normalized to using parameters from the camera calibration to account for distortion in the camera lenses. Camera calibration is determined beforehand and loaded into the GUI for processing. The camera calibration includes information such as focal length, rotation matrix, and translation vector between the two stereo cameras.

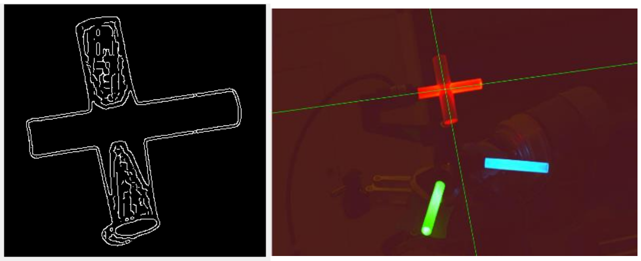

(a)
(b)

Figure 6. Center point identification on red cross

The 2-D to 3-D construction uses two cameras calibrated poses (not detailed in this paper). In each camera, a vector is found from the pinhole camera origin to the cross center. The intersection of the two vectors is found as the location of the red cross in 3D space as shown in Fig. 7.

For both green and blue cylinders from the two images, a plane is found using the origin and the specific line. The line of intersection of the corresponding blue or green planes between the two frames is found. This line of intersection is the centerline through the green or blue cylinder in 3D space.

The position information about the smart target can be displayed on the GUI and will update with each new pair of images from the cameras. The position data is also written to a user-specified file with a timestamp. The next step is to optimize the algorithm to speed up the calculation. To capture dynamic movement, the measurement speed needs to be above 30 frames per second. Adding GPU (Graphics Processing Unit) and parallel calculation will help to speed up the image processing process.

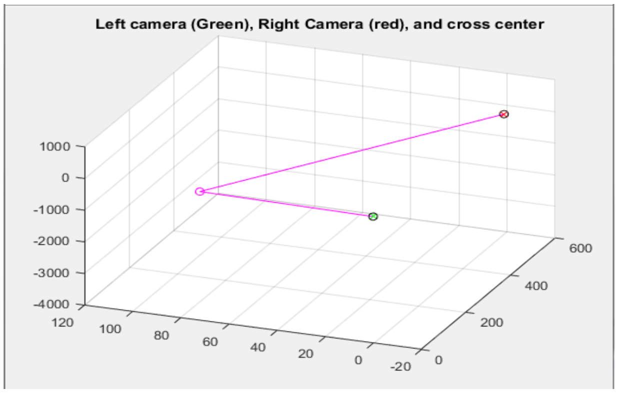


Figure 7. Intersecting vectors from pinhole cameras to center of the red cross

USE CASE DEVELOPMENT

NIST is developing a quick health assessment methodology using a smart target to assess the accuracy degradation of the TCP throughout the robot workspace. The methodology includes the advanced sensing development (the smart target) to acquire the robot TCP's 6-D information, a test method to define the robot movements and a robot error model to reflect the robot geometric and non-geometric errors, and algorithms to process measured data to assess the robot's accuracy degradation.

As shown in Fig. 8 (a), the 6-D smart target is mounted on the last joint of the UR3 robot. A vision-based measurement instrument is set up in the environment, which is at the opposite end of the kinematic chain from the target. The idea is to compare the measured TCP position to nominal positions. A set of predefined robot movements is created (as shown in Fig. 9) based upon the robot's kinematics, geometry, available working volume, and expected operational activities. The left picture of Fig. 9 shows the generation of the target positions. The right picture of Fig. 9 shows the simulation of the robot moving to the

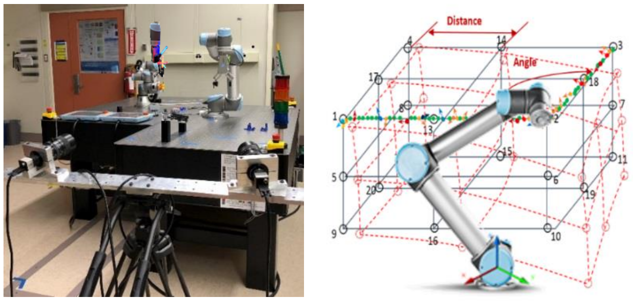
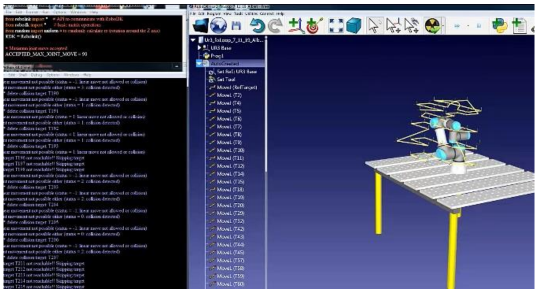


Figure 8. Use case setup for robot quick health assessment



```
from robolink import *    # API to communicate with RoboDK
from robodk import *      # basic matrix operations
from random import uniform # to randomly calculate rz (rotation around the Z axis)
RDK = Robolink()

# Maximum joint move accepted
ACCEPTED_MAX_JOINT_MOVE = 90
```

```
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 3: collision detected)
* delete collision target T190
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 1: collision detected)
* delete collision target T191
ear movement not possible (status = 1: linear move not allowed or collision)
nt movement not possible either (status = 1: collision detected)
* delete collision target T192
ear movement not possible (status = 1: linear move not allowed or collision)
nt movement not possible either (status = 1: collision detected)
* delete collision target T193
ear movement not possible (status = 1: linear move not allowed or collision)
target T196 not reachable!! Skipping target
target T197 not reachable!! Skipping target
target T198 not reachable!! Skipping target
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 2: collision detected)
* delete collision target T203
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 2: collision detected)
* delete collision target T204
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 0: collision detected)
* delete collision target T205
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 0: collision detected)
* delete collision target T206
ear movement not possible (status = -1: linear move not allowed or collision)
nt movement not possible either (status = 2: collision detected)
* delete collision target T207
target T211 not reachable!! Skipping target
target T212 not reachable!! Skipping target
target T213 not reachable!! Skipping target
target T214 not reachable!! Skipping target
target T215 not reachable!! Skipping target
```

Figure 9. Auto-generation of pre-defined robot movement for quick health assessment

planned positions with collision detection algorithms embedded. Unreachable positions or positions having collision problem is removed and updated in the display. The robot movement is measured by the vision-based instrument. The measured TCP 6-D data is used to calculate the deviations from robot normal positions. The calculated deviations are input into the robot error model. The error model handles both the geometric/non-geometric errors and the uncertainties of the measurement system. An algorithm is developed to process the data to assess the robot's accuracy degradation [15]. The first output is the derived error from the calculation of the robot tool center accuracy of the robot through the workspace Fig. 8 (b). The results are more accurate because they are derived from the error model instead of directly calculating from the limited size of sample measurements. This output can be used to view the overall accuracy of a robot within the working volume and find the sweet zone of the robot where the accuracy is suitable for production. The second output is the identified maximum likelihood estimation of axis error parameters. By observing the error pattern, users can monitor the change of robot accuracy. Further analysis can be used to identify the potential error sources of the given errors (for example, zero shift of a joint encoder).

The quick health assessment can be used to swiftly (within 10 minutes) detect degradations in robot accuracy by finding the robot pose deviations from the nominal poses. The use of this methodology will monitor the degradation of robot performance, reduce unexpected shutdowns, and help the optimization of maintenance strategy to improve productivity.

CONCLUSION

Accuracy degradation impacts machine/robot's performance. NIST's development of smart target can be integrated with a vision system to acquire high accuracy position and orientation information of a moving object, allowing for machine/robot's accuracy degradation assessment and accuracy improvement. This paper presented the novel design of the smart target and a software tool. With smart target's measurement, deviations of machine/robot position and orientation are quantified, leading to more accurate calibration or control, improving overall performance. NIST is seeking to develop additional industrial use cases for further applications.

NIST DISCLAIMER

Certain commercial entities, equipment, or materials may be identified in this document in order to illustrate a point or concept. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] A. D. Pham and H. J. Ahn, "High precision reducers for industrial robots driving 4th industrial revolution: state of arts, analysis, design, performance evaluation and perspective," International Journal of Precision Engineering and Manufacturing-Green Technology, vol. 5, pp. 519-533, Aug 2018.

[2] A. Buschhaus, A. Blank, C. Ziegler, and J. Franke, "Highly Efficient Control System Enabling Robot Accuracy Improvement," Procedia CIRP, vol. 23, pp. 200-205, 2014.

[3] A. Cachada, J. Barbosa, P. Leitño, C. A. S. Gcraldcs, et al., "Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture," in 2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA), 2018, pp. 139-146.

[4] D. D. Chen, P. J. Yuan, T. M. Wang, Y. Cai, and L. Xue, "A Compensation Method for Enhancing Aviation Drilling Robot Accuracy Based on Co-Kriging," International Journal of Precision Engineering and Manufacturing, vol. 19, pp. 1133-1142, Aug 2018.

[5] J. Kim, A. Kawamura, Y. Nishioka, and S. Kawamura, "Mechanical design and control of inflatable robotic arms for high positioning accuracy," Advanced Robotics, vol. 32, pp. 89-104, 2018.

[6] D. Culla, J. Gorrotxategi, M. Rodriguez, J. B. Izard, P. E. Herve, and J. Canada, "Full Production Plant Automation in Industry Using Cable Robotics with High Load Capacities and Position Accuracy," in Robot 2017: Third Iberian Robotics Conference, Vol 2. vol. 694, A. Ollero, A. Sanfeliu, L. Montano, N. Lau, and C. Cardeira, Eds., ed Cham: Springer International Publishing Ag, 2018, pp. 3-14.

[7] Q. Guixiu, "Advanced Sensor and Target Development to Support Robot Accuracy Degradation Assessment," in 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), 22-26 Aug. 2019, Piscataway, NJ, USA, 2019, pp. 54-9.

[8] N. Zaimovic-Uzunovic and S. Lemes, "Cylindricity Measurement on a Coordinate Measuring Machine," in Advances in Manufacturing, A. Hamrol, O. Ciszak, S. Legutko, and M. Jurczyk, Eds., ed Cham: Springer International Publishing Ag, 2018, pp. 825-835.

[9] T. F. Shu, S. Gharaaty, W. F. Xie, A. Joubair, and I. A. Bonev, "Dynamic Path Tracking of Industrial Robots

With High Accuracy Using Photogrammetry Sensor," ASME Transactions on Mechatronics, vol. 23, pp. 1159-1170, Jun 2018.

[10] N. Y. Shen, Z. M. Guo, J. Li, L. Tong, and K. Zhu, "A practical method of improving hole position accuracy in the robotic drilling process," International Journal of Advanced Manufacturing Technology, vol. 96, pp. 2973-2987, May 2018.

[11] R. Schares, S. Schmitt, M. Emonts, K. Fischer, R. Moser, and B. Fruhauf, "Improving Accuracy of Robot-Guided 3d Laser Surface Processing by Workpiece Measurement in a Blink," in High-Power Laser Materials Processing: Applications, Diagnostics, and Systems Vii. vol. 10525, S. Kaierle and S. W. Heinemann, Eds., ed Bellingham: SPIE-Int Soc Optical Engineering, 2018.

[12] E. Pivarciova, P. Bozek, Y. Turygin, I. Zajacko, A. Shchenyatsky, S. Vaclav, et al., "Analysis of control and correction options of mobile robot trajectory by an inertial navigation system," International Journal of Advanced Robotic Systems, vol. 15, p. 15, Jan 2018.

[13] H. Gattringer, M. Neubauer, D. Kaserer, and A. Muller, "A Novel Method for Geometric Robot Calibration Using Laser Pointer and Cameras," in Advances in Service and Industrial Robotics. vol. 49, C. Ferraresi and G. Quaglia, Eds., ed Cham: Springer International Publishing Ag, 2018, pp. 200-207.

[14] R. Mautz, "Overview of current indoor positioning systems," Geodesy and Cartography, vol. 35, pp. 18-22, 2009.

[15] G. Qiao and B. A. Weiss, "Industrial Robot Accuracy Degradation Monitoring and Quick Health Assessment," Journal of Manufacturing Science and Engineering, Transactions of the ASME, vol. 141, 2019.

[16] R. Ahmad and P. Plapper, "Safe and Automated Assembly Process using Vision Assisted Robot Manipulator," Procedia CIRP, vol. 41, pp. 771-776, 2016.

[17] A. Filion, A. Joubair, A. S. Tahan, and I. A. Bonev, "Robot calibration using a portable photogrammetry system," Robotics and Computer-Integrated Manufacturing, vol. 49, pp. 77-87, Feb 2018.

[18] Y. B. HuangFu, L. B. Hang, W. S. Cheng, L. Yu, C. W. Shen, J. Wang, et al., "Research on Robot Calibration Based on Laser Tracker," in Mechanism and Machine Science. vol. 408, X. Zhang, N. Wang, and Y. Huang, Eds., ed Singapore: Springer-Verlag Singapore Pte Ltd, 2017, pp. 1475-1488.

[19] M. Ulrich, A. Forstner, and G. Reinhart, "High-accuracy 3D image stitching for robot-based inspection systems," in 2015 IEEE International Conference on Image Processing, ed New York: IEEE, 2015, pp. 1011-1015.

17th World Conference on Earthquake Engineering, 17WCEE

Sendai, Japan - September 13th to 18th 2020

Paper N° C000444 (Abstract ID)

Registration Code: S- A01918

DEVELOPMENT OF A DATABASE OF EXPERIMENTAL TESTS ON FRP RETROFITTED REINFORCED CONCRETE SHEAR WALLS

J. Dukes[1], S. Sattar[2]

[1] *Research Structural Engineer, National Institute of Standards and Technology, jazalyn.dukes@nist.gov*
[2] *Research Structural Engineer, National Institute of Standards and Technology, siamak.sattar@nist.gov*

Abstract

Codes and standards developing organizations, such as American Society of Civil Engineers (ASCE) and American Concrete Institute (ACI), rely on experimental data to develop modeling parameters and acceptance criteria to be used in the performance-based seismic design (PBSD) approach. This data is especially important when developing best practices using materials and techniques that are not covered in current PBSD standards, such as the use of fiber reinforced polymers (FRP) for seismic retrofit. Currently, guidance on the design of a retrofit system for reinforced concrete (RC) shear walls with FRP is missing from standards such as ACI 369.1 and ASCE 41. This paper describes the development of a database of FRP retrofitted shear walls that can assist in filling the information gap in the retrofit design guidance of RC shear walls. The main purpose of this database is to aid in the development of modeling parameters and acceptance criteria that standards developing organizations can adapt in order to provide guidance on the design of FRP retrofitted shear walls. The database includes over 200 experimental tests of shear walls from around the world, which can be filtered into categorized bins, like bins of walls with and without openings, or bins based on the premise of the test (i.e., pre-damaged and repaired vs. undamaged and retrofitted). The database will be available through the DesignSafe-CI platform in the near future, in a format that can be easily accessed and utilized by practicing engineers and researchers. This paper will conclude with recommendations for future experimental tests that would fill in the gaps in the data.

Keywords: retrofit; rehabilitation; fiber reinforced polymer; shear walls; database

Dukes, Jazalyn; Sattar, Siamak. "Development of a Database of Experimental Tests on FRP Retrofitted Reinforced Concrete Shear Walls."
Paper presented at 17th World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Retrofitting of structural components has become an essential technique for engineers and practitioners to reduce the seismic risk associated with existing buildings. The aging of infrastructure and buildings, as well as changing and improved structural codes has led to the need to retrofit many structures, particularly in the United States (U.S.). Retrofitting of reinforced concrete structures and components is increasingly being performed using fiber reinforced polymer (FRP) composites. Application of externally bonded FRP composites onto structures can help repair deteriorated structural components due to degradation or an excessive loading event. Building components can be retrofitted for increased seismic and gravity loads, or to meet current code requirements. Compared with traditional reinforcement materials, FRP composites are light weight and flexible for ease of application, low profile additions to existing structures, and corrosion-resistant. These unique characteristics make FRP composites a desirable retrofit material for existing buildings.

While the benefits of using externally bonded FRP composites to retrofit a structure are known, there is still uncertainty surrounding the initial and long-term performance of these retrofitted components [1-5]. While the use of FRP in structural applications has increased over the past 20 years, entities such as the National Cooperative Highway Research Program (NCHRP) recommend that guidelines, commentary, and examples be developed for design, construction, and maintenance of FRP composites before the material can fully mature and proliferate [6]. One way that engineers and users understand and come to trust the performance of FRP for retrofit of a structure is through experimental testing. There have been many experimental testing programs of FRP retrofitted structural components in the U.S. beginning in the 1980s, including testing of reinforced concrete columns, beams, and beam-column joints [7, 8, 9]. However, experimental research on FRP retrofitted reinforced concrete (RC) shear walls within the U.S. is limited. This lack of available information is concerning, especially as structural walls are continuing to be retrofitted without an equivalent level of knowledge about performance that is available for other components.

This paper describes a database that was developed to fill in the gap of knowledge about available experimental tests that have been performed on FRP-retrofitted reinforced concrete shear walls. There is no other published database that focuses exclusively on FRP-retrofitted reinforced concrete shear walls, to the knowledge of the authors. The database described in this paper details available published research on retrofitted shear walls and compiles them into one location. The database includes over 200 experimental tests of shear walls from around the world, which can be filtered into categorized bins, like bins of walls with and without openings, or bins based on the premise of the test (i.e., pre-damaged and repaired vs. undamaged and retrofitted). This paper details the format and structure of the database, characteristics of the walls studied in the paper, and key research gaps in the database. The intent is to make this database available through the DesignSafe-CI platform [10] in the near future for researchers to use in order to 1) further the understanding of the performance of FRP-retrofitted shear walls, and 2) determine research gaps in the data and develop research programs. Research programs should address gaps in the data that represent the typical walls in existing buildings that are in need of retrofit. Throughout the paper, research gaps and future research needs are noted.

2. Motivation

Several factors motivated the development of this database. First, a workshop held at the National Institute of Standards and Technology (NIST) in 2018 identified the need to understand the extent of experimental testing of FRP-retrofitted structures and components. NIST Special Publication 1244 [11] details the results of the workshop that invited experts and practitioners who use and manufacture FRP materials for design and construction. One of the main concerns identified at the workshop is the lack of large-scale experimental testing of FRP-retrofitted components, and the corresponding lack of understanding of the performance of those structures. From the workshop, as well as a literature review of current research, the idea to focus on

FRP-retrofitted walls was concluded. There is currently no known database that contains all of the information regarding experimental tests performed on FRP-retrofitted walls, so the authors decided to develop the database for future study and for the benefit of the wider community.

Another motivating factor for creating the database is to gather the data needed to enhance current building codes and standards since there are currently no provisions in U.S. design standards such as ACI 369 and ASCE 41 for the design of FRP-retrofitted components. If a designer were designing an FRP-retrofitted wall, for example, there would be no modeling parameters or acceptance criteria that directly relate to that component. The designer may use the backbone curve of an unretrofitted, code-compliant concrete shear wall, but that may not accurately capture the performance of a retrofitted wall. The proposed database will be used to develop modeling parameters and acceptance criteria for inclusion in future iterations of buildings codes and standards. The inclusion of modeling parameters specific to retrofitted components will improve the accuracy of the assessment of the retrofit, and subsequently increase the confidence of the designer in their design of retrofitted components.

3. Database Development

The proposed database contains over 200 reinforced concrete wall tests collected from almost 40 research programs reported in the literature. In the database, each specimen is given a unique specimen identification number (ID) and also retains the specimen name given in the research paper. The digital object identifier (DOI) number that links to the publication from which details about the experimental testing were retrieved is noted in the database, when available. A citation of the paper is also included.

Fig. 1 illustrates the countries of origin of the research programs that are included in the database. Canada has contributed the largest number of research programs of any country to this database. European countries, such as France and Greece, contribute about 30 % of the research programs. Asian countries, such as Singapore and Japan, contribute about 35 % of the research programs. In this database, only 3 of the research programs originate out of institutions based in the U.S. This information suggests that most of the research on the performance on FRP-retrofitted shear walls were on specimens that may not have been designed to the U.S. standards and codes or designed with similar material properties. This is a research gap that should be addressed: to increase the number of experimental tests on FRP-retrofitted shear walls that are designed to prior editions of U.S. codes in order to represent typical shear wall construction in existing structures within the U.S.

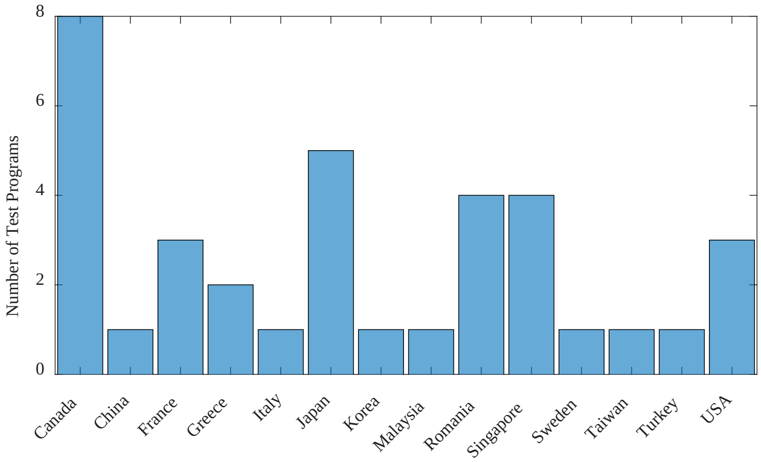


Fig. 1 – Country of origin of all research programs in the database

Experimental research on FRP-retrofitted shear walls began in the late 1990s, according to the research programs included in the database. Approximately ten percent of the papers in the database were published before 2000. About 30 % of papers were published between 2000 and 2009. Almost 60 % of the papers were published in the last decade. This could suggest that interest in determining the performance of FRP-retrofitted walls has been increasing in recent years, and that there may be areas of research yet to be explored.

Wall specimens are categorized in several ways in the database. Two distinctions among the wall specimens were quickly apparent after collecting the information: the presence of openings, and the type of retrofit application. The wall specimens tested in the database either had openings that were built into the wall or created after the solid wall was built, or no openings. The presence of openings in the walls is denoted by a "Y or N" entry, "Y" meaning openings are present, "N" meaning the specimen is a solid wall. The other distinction is related to the presence of damage and repair before the application of the FRP retrofit. For example, many programs built plain (unretrofitted) RC walls, tested them under cyclic loads, repaired then applied FRP and retested the specimen again under a cyclic loading protocol. These specimens are denoted as "Repair" specimens in the database. The other RC walls which were built in the lab and immediately treated with an FRP retrofit without prior testing are denoted as "Retrofit" specimens. The walls that acted as control walls without any FRP retrofit are denoted as "Control". By noting these important distinctions, users of the database can filter out the specimens based on their research needs. Table 1 shows subsets of walls from the database based on these two distinctions. Details about Subsets A and B will be referred to later in this paper, and will also be referred to as Retrofit and Repair subsets, respectively. Subsets A and B represent about half of the total number of specimens in the database and are discussed in more detail in this paper. Details about Subsets C and D are not discussed in this paper.

Table 1 – Subsets of database based on presence of openings and prior damage

Wall Test and Condition	Retrofit No damage prior to FRP		Repair and Retrofit Damage prior to FRP	
No Openings	Subset A Retrofit, no openings	62 specimens	Subset B Repair, no openings	54 specimens
Openings	Subset C Retrofit, with openings	55 Specimens	Subset D Repair, with openings	33 Specimens

4. Characteristics of the walls

4.1 Wall parameters

The data and details stored in the database describe the types of walls that have been tested in the literature. These details include wall parameters such as geometric and material properties, FRP design and material properties, the loading types and testing protocols, and the results from the test including maximum drift ratio and maximum lateral force and the backbone curves. Table 2 shows the ranges of the following wall parameters for the Retrofit and Repair subsets: concrete compressive strength (f'_c), wall aspect ratio (a/l_w), yield strength of the longitudinal steel reinforcement (f_y), horizontal steel reinforcement ratio (ρ_t), longitudinal steel reinforcement ratio (ρ_l), and axial load ratio ($P/A_g*f'_c$). Figs. 2 and 3 show four wall characteristics in the form of histograms for Subsets A and B of the database. By viewing the data in this way, gaps in representation of wall or test conditions become apparent, and wall characteristics which have been tested more than others or have not been tested at all can be identified. For example, most of the tested walls in the database were tested without any axial load (i.e., axial load ratio equals zero), and the highest axial load ratio was less than 10 % for most of the tests in Subsets A and B. Knowing this limitation in available data can direct future research programs, as testing FRP-retrofitted walls under higher load ratios could be of interest for researchers or designers. The same analysis can be performed for the wall aspect ratio, longitudinal steel reinforcement ratio, and horizontal steel reinforcement ratio.

Table 2 – Summary of wall parameters for Subsets A and B

Subset	f'_c (MPa)		a/l_w		f_y (MPa)		ρ_t, %		ρ_l, %		$P/(A_g{*}f'_c)$	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Retrofit (A)	14	45	0.44	3.12	235	500	0.09	0.57	0.16	1.83	0	0.09
Repair (B)	16.6	42	0.85	2.5	320	585	0.25	0.57	0.28	3.0	0	0.2

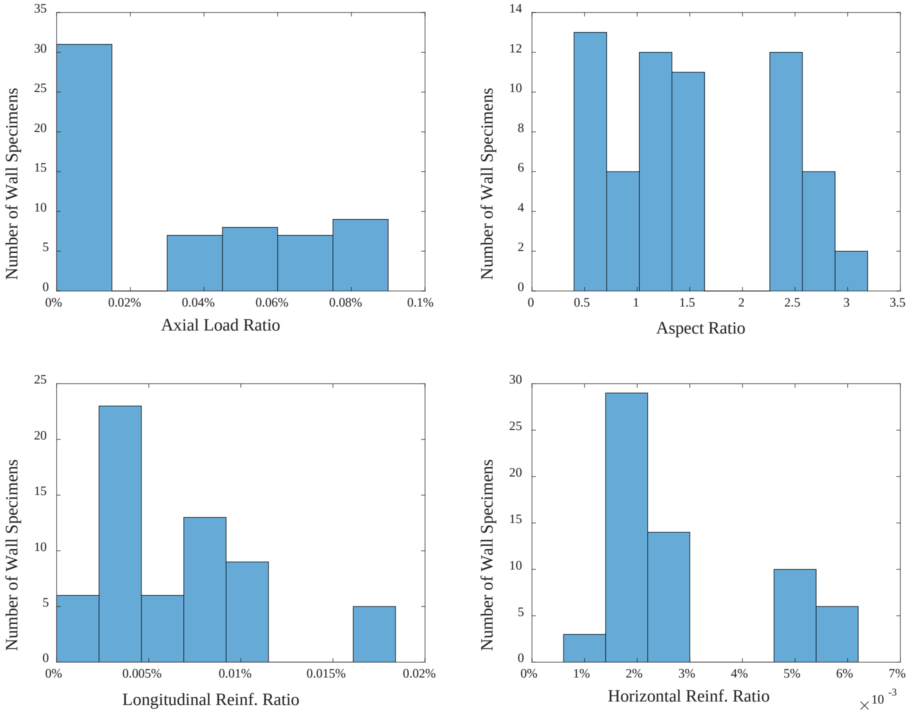


Figure 2 – Histograms of wall parameters of Subset A (Retrofit)

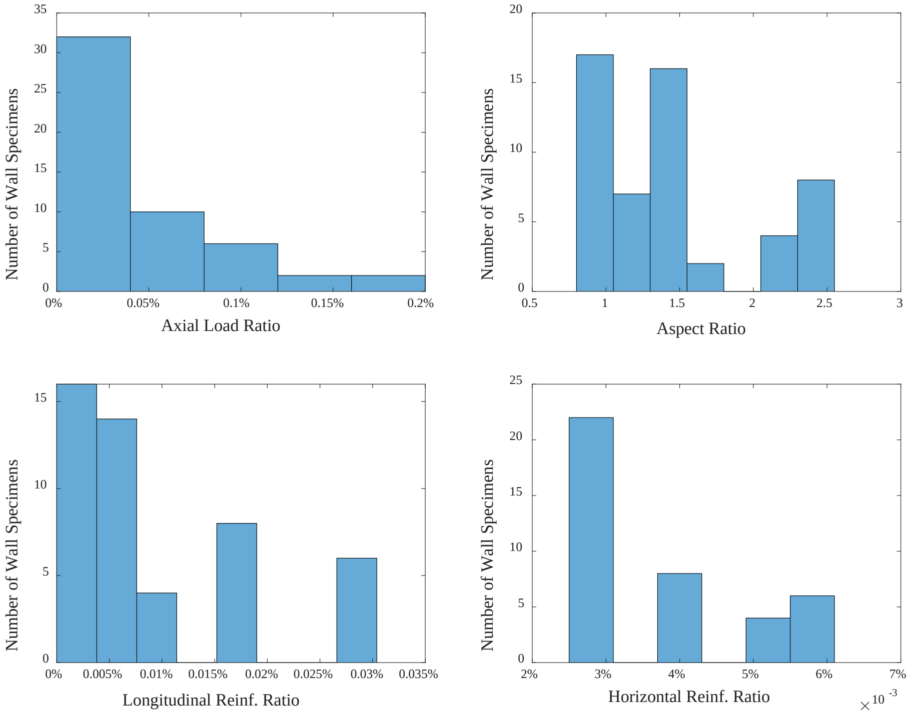


Figure 3 – Histograms of wall parameters of Subset B (Repair)

All of the walls have rectangular cross sections (no L, C or T shaped sections). Most of the walls in the Retrofit subset were loaded with cyclic lateral loading, while a few (~ 12 % of walls) were tested under monotonic lateral loading. All walls in the Repair subset were tested under cyclic lateral loads except for one experimental program that tested walls on a shake table. Less than ten percent of the walls in Subsets A and B were tested to the point of loss of axial load-carrying capacity or significant loss of lateral strength. However, there is value in testing structural components to significant strength loss to observe the behavior of the component as the strength degrades. This information is useful for developing nonlinear models capable of capturing the response of the components in post-peak range of the response. The testing of FRP-retrofitted walls to a state of significant strength loss is an area of research that should be explored. Another area of research to explore is various shapes of walls, including L, C, or T shapes, since these shapes present challenges related to FRP retrofit application as well as a difference in behavior from rectangular walls. This may be a consideration for researchers in the future when creating testing programs.

4.2 Wall design and test objective

The initial wall design of the test specimens in the database varied based on two main objectives found in the test programs featured in the database: shear strengthening or flexural strengthening of the wall. Fig. 4 illustrates the difference in the objective of strengthening that is applied to shear walls. Because FRP is being used as a retrofit technique, it is appropriate that most of the walls were designed to have deficiencies. Several research programs describe their walls to have been designed to older design codes that are now known to have produced walls that perform poorly under seismic loads. Other walls were designed to be shear deficient, or under-reinforced in order to fail in shear. Walls that were designed with built-in deficiencies in shear were then retrofitted with FRP to improve shear strength and ductility of the walls and to determine the effectiveness of the retrofit system. Another group of wall specimens were designed to measure the effectiveness of FRP to enhance the flexural strength of a shear wall. These walls were designed to behave in a ductile manner before the expected shear strength was reached. Still other wall specimens were designed for the FRP to improve both shear and flexural strengthening. This information is important to researchers in determining the past research as well as areas of research that have not been explored in terms of wall design and test objectives.

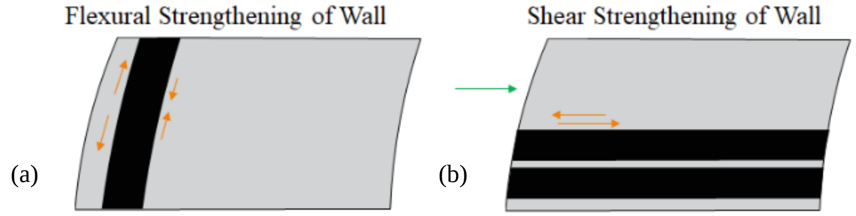


Fig. 4 – Types of FRP strengthening for walls (a) flexural strengthening, and (b) shear strengthening

4.3 FRP retrofit design details

The design of the FRP retrofit varied between test programs, and even between specimens within a test program. No two FRP retrofit designs were the same. Some of the differences between designs include FRP material, number of layers, thickness of the layers, use and spacing of horizontal and vertical laminates, and use of anchors. Fig. 5 shows the types of FRP material used in the walls of subset A and B. Carbon FRP (CFRP) is the most common type of material used, followed by glass FRP (GFRP). Some walls included two or more types of FRP material in the design (designated as “combo”). The focus on CFRP and GFRP in the testing programs are appropriate since these FRP materials are the most common materials used in the U.S. However, if one wanted to determine the performance of a wall retrofitted with FRP material other than carbon or glass, this database does not contain many examples of the performance of alternative materials. This may be a consideration for researchers in the future when creating testing programs.

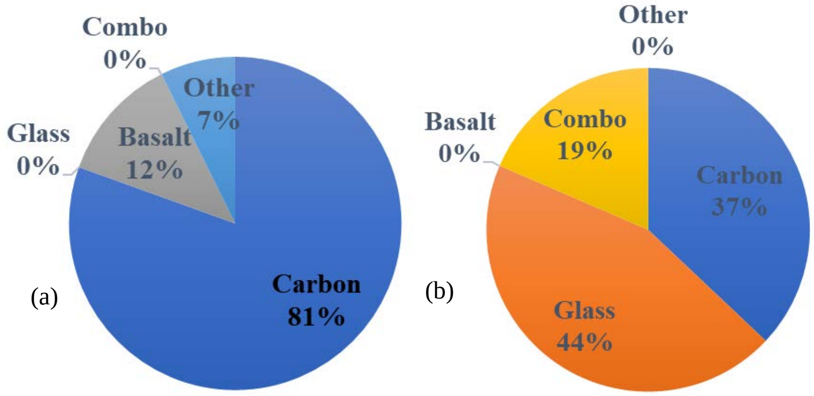


Fig. 5 – Pie chart of FRP material types for (a) Subset A (Retrofit), and (b) Subset B (Repair).

The number of layers and the orientation of FRP layers also varies widely among the tested walls. Walls were either reinforced on one side or both sides. While reinforcing a wall on both sides may provide better confinement and better performance, sometimes it is only practical to retrofit one side of a wall in the field due to limited access to both sides of the wall. Many of the walls included FRP anchors to prevent premature debonding of the FRP from the concrete substrate during testing. FRP anchors are an important part of an FRP retrofit design and are becoming more prominent in the field. However, there are currently no provisions in U.S. design standards on how to design the FRP anchors. Tests that include FRP anchors in the design are helpful for the researchers and practitioners that develop codes and standards to create guidelines related to the FRP anchor design. Fig. 6 illustrates the portion of wall specimens that had FRP anchors and the type of anchors for Subsets A and B.

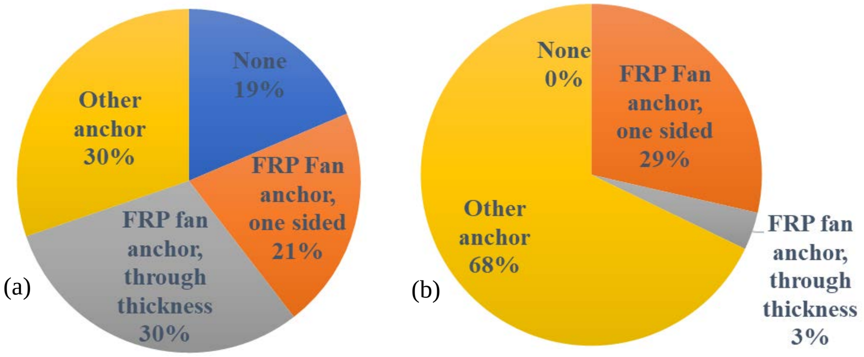


Fig. 6 – Pie chart of FRP anchor types for (a) Subset A (Retrofit), and (b) Subset B (Repair).

5. Conclusion

This study develops a database of experimental tests that have been performed on FRP-retrofitted reinforced concrete shear walls. This database includes all experimental research on FRP-retrofitted walls known to the authors at the point of publication. This database will be published on the DesignSafe-CI platform for future use by researchers to further the understanding of the performance of FRP-retrofitted RC walls. It will also be available for review by researchers who have experimental data to contribute, who can submit their data to the authors for future iterations of the database. In this way, the database will have the most up-to-date collection of experimental data on FRP-retrofitted walls.

This database also helps identify gaps in the existing research, which may lead to future experimental testing programs. One gap that was apparent was the lack of tests that had an axial load ratio above ten percent. About a third of the tested specimens had no axial load applied during testing. Another gap in the database is the lack of tests that tested the specimen to significant lateral strength loss. There is a lot of information to be learned from testing components to collapse or significant lateral strength loss, especially for those interested in capturing the response all the way to reaching the residual strength. It is important to understand the post peak range of the response of FRP retrofitted walls as the wall can potentially experience a brittle mode of failure, i.e., abrupt loss of capacity, when the failure is initiated. Finally, research that investigates the effect of FRP anchors on the performance of the wall specimen is important to practitioners in the field in order to effectively design an FRP retrofit system with anchors. The current standards currently do not provide guidance on the use of anchors for designing the FRP retrofit system for walls. Practitioners need guidance on how to design these anchors, including the spacing, shape, and angle of the anchors in order to achieve the desired performance of retrofitted walls and prevent premature failure/debonding of the FRP retrofit system. A structured research program that systematically varies these parameters on shear walls would be useful in the development of design guidance of FRP anchors.

6. Disclaimer

Commercial software may have been used in the preparation of information contributing to this paper. Identification in this paper is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that such software is necessarily the best available for the purpose. No formal investigation of uncertainty or error is included in this study.

7. References

[1] ACI Committee 440 (2015): 440.9R-15 Guide to accelerated conditioning protocols for durability assessment of internal and external fiber-reinforced polymer FRP reinforcement. *American Concrete Institute*, Farmington Hills, MI.

[2] ACI Committee 440 (2017): 440.2R-17 Guide for the design and construction of externally bonded FRP systems for strengthening concrete structures. *American Concrete Institute*, Farmington Hills, MI.

[3] AC125 ICC-ES (2007): Acceptance Criteria for Concrete and Reinforced and Unreinforced Masonry Strengthening Using Fiber-Reinforced Polymer (FRP) Composite Systems*. International Code Council (ICC)-Evaluation Service*, Whittier, CA.

[4] AC434 ICC-ES (2016): Proposed acceptance criteria for masonry and concrete strengthening using fiber-reinforced cementitious matrix (FRCM) composite systems*. International Code Council (ICC)-Evaluation Service*, Whittier, CA.

[5] AASHTO (2012): Guide specifications for design of bonded FRP systems for repair and strengthening of concrete bridge elements. *AASHTO*, Washington, DC, 1st Ed.

[6] O'Connor, J.S. and Frankhauser, W. (2016): Advances in Fiber-Reinforced Polymer Composites in Transportation Infrastructure. *Transportation Research Record: Journal of the Transportation Research Board* (2592):56-64.

[7] Saadatmanesh, H., Ehsani, M.R., and Li, M.-W. (1994): Strength and ductility of concrete columns externally reinforced with fiber composite straps. *ACI Structural Journal* 91(4):434-447.

[8] Pendhari, S.S., Kant, T., and Desai, Y.M. (2008): Application of polymer composites in civil construction: A general review. *Composite Structures* 84(2):114-124.

[9] Alvarez, J.C., Brena, S.F., and Arwade, S.R. (2018): Nonlinear Backbone Modeling of Concrete Columns Retrofitted with Fiber-Reinforced Polymer or Steel Jackets. *ACI Structural Journal*, V. 115, No. 1, January 2018.

[10] Rathje, E., Dawson, C. Padgett, J.E., Pinelli, J.-P., Stanzione, D., Adair, A., Arduino, P., Brandenberg, S.J., Cockerill, T., Dey, C., Esteva, M., Haan, Jr., F.L., Hanlon, M., Kareem, A., Lowes, L., Mock, S., and Mosqueda, G. (2017): DesignSafe: A New Cyberinfrastructure for Natural Hazards Engineering. *ASCE Natural Hazards Review*, doi:10.1061/(ASCE)NH.1527-6996.0000246. https://www.designsafe-ci.org/

[11] Goodwin, D., Sattar, S., Dukes, J., Kim, J.K., Ferraris, C., and Sung, L. (2019): Research Needs Concerning the Performance of Fiber Reinforced (FR) Composite Retrofit Systems for Buildings and Infrastructure. *NIST Special Publication 1244*, December, 2019.

A Cryogenic Quantum-Based RF Source

J. A. Brevik[1,#], A. S. Boaventura[1,2], M. A. Castellanos-Beltran[1], C. A. Donnelly[1], N. E. Flowers-Jacobs[1], A. E. Fox[1], P. F. Hopkins[1], P. D. Dresselhaus[1], D. F. Williams[1], S. P. Benz[1]
[1]National Institute of Standards and Technology, Boulder, CO 80305 USA
[2]University of Colorado, Boulder, CO 80305 USA
[#]justus.brevik@nist.gov

***Abstract*— We performed a preliminary calibrated measurement of the output power of a Josephson arbitrary waveform synthesizer up to 1 GHz. We present the results and measurement procedure for generating quantum-based signals using an array of Josephson junctions operating at cryogenic temperature and calibrating those signals to transfer the on-wafer quantum-based accuracy to room temperature.**

***Keywords*— Josephson arrays, quantization, signal synthesis, standards, superconducting integrated circuits, voltage measurement, power measurement, digital-analog conversion.**

I. Introduction

The Josephson arbitrary waveform synthesizer (JAWS) has been demonstrated to be an invaluable tool for synthesizing quantum-based voltage signals for metrology in the audio-frequency range [1]. Our goal is to extend the metrology capability of the JAWS system as a quantum-based source to the radio-frequency range. Our initial effort has extended the maximum waveform synthesis frequency to several gigahertz, while our goal is to eventually increase it to tens then hundreds of gigahertz [2].

The synthesis of arbitrary waveforms using pulse-driven Josephson junctions (JJs) is detailed in [1],[3]. Briefly, waveforms are encoded into a pattern of return-to-zero pulses with two or more levels using a delta-sigma modulation algorithm that modulates the separation of the pulses. The output patterns have calculable spectra with high spurious-free dynamic range (SFDR) and signal-to-noise ratio (SNR) near the fundamental synthesis frequency, despite the large digitization error resulting from the limited number of encoded pulse levels. When this pulse pattern is generated by a standard room-temperature digital-to-analog converter (DAC), the output signal will fluctuate with time and environmental conditions and will have degraded SFDR and SNR compared to the calculated spectrum. The signal from this DAC output can be conditioned by an array of JJs, which acts as a pulse quantizer. The resulting signal is stable, has an amplitude with quantum-based accuracy, and a spectrum that has exceptionally high SFDR and SNR.

Under the proper conditions, when a JJ is biased using a current pulse it will generate a quantized voltage pulse with an integrated time area that is exactly equal to a multiple of a magnetic flux quantum $\Phi_0 = h/2e$, where h is the Planck constant and e is the elementary unit of electric charge. Through this effect, an array of JJs can be used to generate quantized output pulses that have stable and invariant time-integrated area based on quantum effects and tied to fundamental constants.

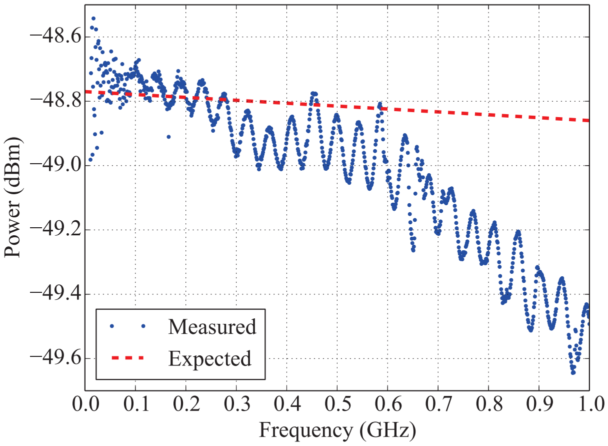


Fig. 1. The calibrated output power versus synthesis frequency for the JAWS system (dots) and the expected frequency-dependent value (dashed line).

Unique to JJ circuits, the quantized pulses are generated over a range of bias current amplitude and other conditions called the quantum locking range (QLR). When the JAWS system operates within its QLR, every incident bias pulse is transformed by every JJ in the array into a quantized output voltage pulse. This ensures that the output amplitudes of the synthesized waveforms are quantum-based, stable, independent of operating location, and immune to small variations in operating bias, temperature and other conditions.

JAWS systems have synthesized waveforms up to 1 GHz, but those signals were not corrected for the errors occurring between the JJ devices and the room-temperature measurement [4]. In this work, we used a cryogenic probe station to measure the output signals of a JAWS circuit and a set of cryogenic calibration standards to reference the signals measured at room temperature to the signals generated on chip, where they have quantum-based accuracy [5],[6].

II. Experimental Setup

As a first step in demonstrating the quantum-based output of the JAWS system, we have measured and corrected its output power versus synthesis frequency (frequency response) up to 1 GHz, as shown in Fig. 1. The goal of the experiment was to demonstrate that the frequency response could be

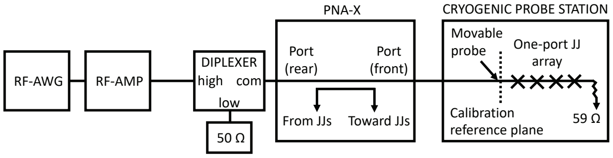


Fig. 2. Block diagram of the measurement setup.

generated and properly corrected so that it is in agreement with the expected value and independent of the synthesis frequency, within the limits discussed in section III-C. Calibrated phase measurements of the JAWS output signals will be included in future experiments.

A schematic representation of the measurement setup is shown in Fig. 2. The main equipment used in the experiment includes a cryogenic probe station with a cryogenic chip containing the JJ circuit and standards, a Keysight N5242A precision network analyzer (PNA-X), and a Keysight M8195A 65 gigasample-per-second arbitrary waveform generator (RF-AWG)[1]. The RF-AWG was used to apply the current bias pulses to the JJ array, and an RF amplifier was used to increase their amplitude. The bias signal was transmitted to the high-frequency port of a diplexer with a 5 GHz crossover frequency. The low-frequency port of the diplexer was terminated with a 50 Ω resistor, and the common port was connected to the rear-panel input to one port of the PNA-X. The front panel of that port was connected to a movable cryogenic microwave probe via a feedthrough port on the probe station.

The current bias pulse signal contains a component at the fundamental synthesis frequency that drives the distributed inductance of the JJ array coplanar waveguide (CPW), creating a frequency-dependent voltage error that adds vectorially to the desired signal [3]. The diplexer was used to high-pass filter the current bias pulses and reduce this so-called "feedthrough" error. However, the high-pass port presented a high impedance to the low-frequency synthesized (<5 GHz) signal traveling from the JJ array toward the PNA-X. To limit the reflection of the synthesized signal from the high-pass port, a diplexer was used instead of a high-pass filter so that a $50\ \Omega$ termination was presented to the JJ output signal.

We used a PNA-X to measure the signals from the JJ array and to measure the calibration standards. Only one port of the PNA-X was used for the JJ device measurement, but two ports were used during the on-chip calibration standard measurements. The PNA-X port that was connected to the JJ circuit includes an internal bias tee, which was used to apply an additional dc bias current to the bias pulse pattern when the QLR was measured. A comb generator, triggered by a 10 MHz square wave generated by a synchronized channel of the RF-AWG, was used as a phase reference for the PNA-X.

For this first experimental iteration, we measured a one-port JAWS circuit, as shown in Fig. 2. In this circuit, the JJs are embedded in a CPW in 500 vertical stacks with three JJs per stack, for a total of 1500 junctions [7]. One end of the CPW is connected to the ground plane via a termination resistor with a design impedance of $50\ \Omega$ to absorb the residual bias pulse signal. The opposite end of the CPW has contact pads to accept a microwave probe that was used to apply the bias signals and measure the JJ output signal. This circuit topology provides roughly half of the full signal of the quantized output pulses generated by the JJ array, as explained in section III-C.

A set of calibration standards were fabricated in niobium on the same chip as the JAWS circuit. These superconducting calibration standards provide the basis for the second-tier calibration of the PNA-X down to the wafer reference plane. They include a multi-line ("thru"-reflect-line) TRL kit with five line standards ranging in length from $70\ \mu\mathrm{m}$ to $9\ \mathrm{mm}$, and several short, thru, open, and load standards. Additional details about the calibration kit can be found in [6].

We used a cryogenic probe station to measure the signals generated by the JAWS circuit and to perform the cryogenic calibration of those signals. The chip containing the cryogenic standards and JJ device was mounted on a cryogenic stage that was controlled at 4 K. The cryogenic probe station has two microwave probes that can each be controlled along three axes of motion using cryogenic piezoelectric nanopositioners. The microwave probes were used to measure the complex voltage waveforms traveling toward and away from the device under test, and one probe was also used to apply the bias signals to the JJ array. An optical microscope was used to image the 4 K stage through transparent windows in the cryostat jacket and radiation shields to position the probes on the cryogenic chip.

III. Measurement

A. Measurement Procedure

We measured the frequency response of the JAWS output signal by measuring the power at the fundamental of single output tones that were synthesized in steps of 1 MHz from 10 MHz to 1 GHz. These 991 single-tone bipolar waveforms were each encoded into a bias pulse pattern with a minimum of $10\,000$ waveform periods using a second-order, three-level $[-1, 0, +1]$, bandpass delta-sigma modulator at a 64 gigasample-per-second rate. To reduce the signal at the fundamental in the bias pulse pattern and decrease the feedthrough error, the three-level codes were transformed to five-level codes so that each bias pulse had bracketing half-amplitude pulses and each pulse block had effectively zero average current [3]. For this experiment, quantum-locked operation could be achieved across the entire synthesis frequency range only by reducing the amplitude of the waveforms to 2.5 % of the full scale possible with the given sample rate. The limitation on the mark density of the pulses is under investigation, but an increase in the signal amplitude should be possible in future measurements.

[1] Commercial instruments are identified in order to adequately specify the experimental procedure. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the equipment identified is necessarily the best available for the purpose.

We performed the measurement by sequentially programming each of the 991 pulse patterns into the RF-AWG, setting the PNA-X measurement frequency to the corresponding value of the synthesis frequency, and measuring the forward and backward complex waveforms at the input of the JJ circuit. During these measurements the dc offset bias was set to zero and the cryostat windows were closed.

B. Calibration

After measuring the JAWS signals, we performed the calibration measurements that were used to correct the raw frequency response data. The calibration procedure is described in detail in [5],[6], but a brief description of the procedure follows. We measured the on-chip cryogenic standards on a 991-point frequency grid matching the single-tone synthesis values. To establish the second-tier calibration reference plane at the end of the microwave probes, we used both probes and PNA-X ports to measure the two-port cryogenic TRL standards. Next, we disconnected the PNA-X coaxial cables from the probe station and performed a short-open-load-thru (SOLT) calibration at the end of the cables to determine the first-tier calibration reference plane. We then calibrated the absolute amplitude using a power meter that is traceable to the NIST calorimeter power reference, and calibrated the absolute phase using a frequency comb generator that is traceable to the NIST electro-optic sampling system. We used the NIST microwave uncertainty framework (MUF) software to compute the error correction [8], which we then used to correct the JJ signal measurements to produce the frequency response shown in Fig. 1. The uncertainty for the correction has not yet been calculated.

C. Expected Signal Level

When generating quantum-based JAWS signals, the exact dc value of the quantized pulse area across the JJ array is calculable. There are several additional effects that must be considered when calculating the expected output signal level, especially at higher frequencies, that are discussed below. The calculable dc signal level—assuming the pulses are ideal delta functions and synchronized—depends on the number of JJs in the array, the maximum density of pulses, and the number of quantized pulses generated for each input bias pulse. The codes used in this experiment were all designed for a -42 dBm output signal.

Because we measured a one-port JJ device in this experiment, only a portion of each quantized output pulse area was collected and an additional correction to the calculable signal level is required. As detailed in [9], the quantized output pulses split with a portion propagating along the CPW in the direction of the bias pulses and the other portion traveling in the opposite direction. The ratio of the split is determined by the voltage division of the impedances seen in either direction from the point of generation. The measured dc impedance of the nominally 50 Ω termination resistor was 59 Ω, and the measured dc impedance at the probe looking toward the PNA-X was 50 Ω. Therefore, approximately 46 % of the quantized pulse area was measured, reducing the total power by -6.8 dB to a total expected dc power level of ~ -48.8 dBm.

The quantized pulses are not perfect delta functions but have a finite pulse width that causes the signal level to decrease at higher synthesis frequencies. There are two main contributions to the finite pulse width. The first is simply due to the finite characteristic frequency (~ 20 GHz) of our JJs, which we expect to reduce the signal by -0.06 dB at 1 GHz [4]. The second effect relates to pulse propagation. Due to the finite distribution of the JJs along the CPW, the sum of the quantized pulses that propagate in the same direction as the incident bias signal has a narrower overall pulse shape than the summed pulse traveling in the opposite direction. Since a one-port circuit was used in this work, we measured the broader composite pulses. Given 500 JJ stacks at the standard 6.5 μm stack spacing, the estimated output power reduction is -0.8 % or -0.03 dB at 1 GHz [9]. These two pulse width effects have negligible impact on the total expected power at 10 MHz, but reduce it by -0.09 dB to around -48.9 dBm at 1 GHz. There is a corresponding negative slope in the expected frequency response between 10 MHz and 1 GHz of -0.09 dB, which is a 2.1 % error in power. This error can be reduced in the future by increasing the characteristic frequency of the JJs.

In previous measurements, the contribution of the feedthrough error was considerable, especially at 1 GHz [4]. In this experiment, the feedthrough was effectively eliminated by the improved filtering provided by the stop-band of the high-pass diplexer filter. The feedthrough error was reduced below the noise floor of the PNA-X for even a 1 Hz measurement bandwidth, setting its upper limit at -120 dBm at 1 GHz.

D. Verification of Quantum Locking Range

It is critical that the system operates within the QLR of the JJ circuit at each synthesis frequency step to ensure that the output power at the fundamental synthesis tone is quantum-based and constant, within the limits described in the previous section. We verified the QLR for each of the 991 synthesis steps by measuring the JAWS output with respect to a dc offset current applied to the JJ array. As the magnitude of the dc offset increases from zero, the JJs will eventually either fail to emit a quantized output pulse for each bias pulse or will emit more quantized pulses than intended, thus defining the boundaries of the QLR. When this boundary is crossed, there is little change in the output power at the fundamental. However, any omission or addition of pulses in the quantized pattern will cause the noise floor of the calculable output spectrum near the fundamental to change in an appreciable way.

We measured the QLR by stepping an offset current bias applied to the array and recording the power at the 50 noise spurs above and the 50 spurs below the fundamental synthesis tone that make up the calculable noise floor. The PNA-X was configured for 10 Hz bandwidth and 101 measurement points on a frequency grid that captured the synthesis tone and the 100 nearest noise spurs. The power was recorded for each

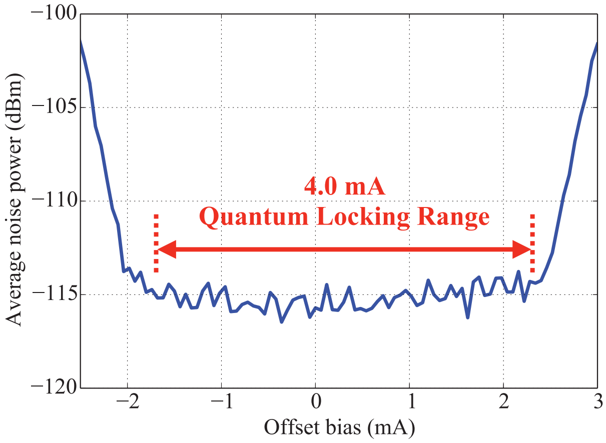


Fig. 3. Characteristic measurement of the quantum locking range (QLR) with respect to offset bias current for a single synthesis frequency step (300 MHz). The average power of the noise floor near the synthesis tone is plotted against the offset bias current. The dashed lines indicate the boundaries of the 4.0 mA QLR measured for this synthesis frequency step.

of the 101 frequency points while stepping the offset bias from -3 mA to +3 mA. The average of the power for the 100 noise spurs was then plotted against the offset bias as in Fig. 3, which shows a characteristic QLR measurement for a single-tone synthesis step. An inflection point occurs where the average noise power starts to increase with the bias offset, corresponding to the boundaries of the QLR. The span in current from the negative to positive inflection points was calculated for each synthesis frequency. We measured a minimum QLR of 3.8 mA across the entire synthesis range, demonstrating the output of the JAWS source was measured while in quantum-locked operation.

E. *Calibrated Frequency Response*

The resulting calibrated frequency response is shown in Fig. 1, along with the calculated frequency-dependent power from section III-C. At low frequency there is reasonable agreement between the measured and expected values. However, as the synthesis frequency increases, the measured value drops further below the expected value and we observe an excess frequency-dependent "roll-off." In addition, the response shows a ripple with roughly 0.2 dB peak-to-peak amplitude. The maximum deviation in the measured power from the expected frequency-dependent value occurs at 967 MHz, where the measured power is -49.6 dBm. This error is equivalent to about -0.8 dB or 17 % in power.

The source of the additional roll-off is under investigation. It is possible that we have underestimated the width of the quantized output pulses and therefore have underestimated the expected roll-off. We plan to make calibrated measurements of the widths of the input bias and quantized output pulses at the wafer reference plane to re-evaluate our estimate. The origin of the ripple is also under investigation, and we should be able to remove it in future measurements.

IV. Conclusion

We have produced a preliminary calibrated measurement of the frequency response of the output power of a JAWS circuit up to 1 GHz. The calibrated output power differs by a maximum of -0.8 dB, or 17 % in power, from the expected value. This large deviation is due to a combination of ripple and excess roll-off of the signal at higher synthesis frequencies. The latter contribution is the dominant source of error, and we attribute it to a larger than expected quantized pulse width. This will be mitigated in the future by optimizing the bias pulses and by increasing the characteristic frequency of the JJs. The output power is currently about -49 dBm, but we should be able to increase this to -30 dBm, which is the level where the source becomes most useful for RF metrology. We can do so by transitioning to two-port device measurements, by increasing the encoded signal amplitudes, and by increasing the number of JJs in the array. Finally, we plan on estimating the uncertainty of the calibration procedure in our next measurement campaign and including it in our error analysis.

Acknowledgment

The authors thank the NIST Boulder Microfabrication Facility. The authors also thank Adam Sirois for his work on the cryogenic probe station. This research was supported by NIST's Innovations in Measurement Science program. This work is a contribution of the U.S. Government and is not subject to U.S. copyright.

References

[1] S. P. Benz and C. A. Hamilton, “A pulse-driven programmable Josephson voltage standard,” *Applied Physics Letters*, vol. 68, no. 22, pp. 3171–3173, 1996.

[2] P. F. Hopkins, J. A. Brevik, M. Castellanos-Beltran, C. A. Donnelly, N. E. Flowers-Jacobs, A. E. Fox, D. Olaya, P. D. Dresselhaus, and S. P. Benz, “RF waveform synthesizers with quantum-based voltage accuracy for communications metrology,” *IEEE Transactions on Applied Superconductivity*, vol. 29, no. 5, pp. 1–5, Aug 2019.

[3] J. A. Brevik, N. E. Flowers-Jacobs, A. E. Fox, E. B. Golden, P. D. Dresselhaus, and S. P. Benz, “Josephson arbitrary waveform synthesis with multilevel pulse biasing,” *IEEE Transactions on Applied Superconductivity*, vol. 27, no. 3, pp. 1–7, April 2017.

[4] C. A. Donnelly, N. E. Flowers-Jacobs, J. A. Brevik, A. E. Fox, P. D. Dresselhaus, P. F. Hopkins, and S. P. Benz, “1 GHz waveform synthesis with Josephson junction arrays,” *IEEE Transactions on Applied Superconductivity*, vol. 30, no. 3, pp. 1–11, April 2020.

[5] A. S. Boaventura, J. A. Brevik, D. F. Williams, A. E. Fox, P. F. Hopkins, P. D. Dresselhaus, and S. P. Benz, “Calibrating a quantum-based radio-frequency source,” *ARFTG 2020*, submitted for publication.

[6] A. S. Boaventura, D. F. Williams, R. A. Chamberlin, J. G. Cheron, A. E. Fox, P. D. Dresselhaus, P. F. Hopkins, I. W. Haygood, and S. P. Benz, “Microwave modeling and characterization of superconductive circuits for quantum voltage standard applications at 4 K,” *IEEE Transactions on Applied Superconductivity*, vol. 30, no. 2, pp. 1–7, March 2020.

[7] B. Baek, P. D. Dresselhaus, and S. P. Benz, “Co-sputtered amorphous Nb$_x$ Si$_{1-x}$ barriers for Josephson-junction circuits,” *IEEE Transactions on Applied Superconductivity*, vol. 16, no. 4, pp. 1966–1970, Dec 2006.

[8] www.nist.gov/services-resources/software/wafer-calibration-software.

[9] C. A. Donnelly, J. A. Brevik, N. E. Flowers-Jacobs, A. E. Fox, P. D. Dresselhaus, P. F. Hopkins, and S. P. Benz, “Quantized pulse propagation in Josephson junction arrays,” *IEEE Transactions on Applied Superconductivity*, vol. 30, no. 3, pp. 1–8, April 2020.

Evolving Advanced Persistent Threat Detection using Provenance Graph and Metric Learning

Gbadebo Ayoade*, Khandakar Ashrafi Akbar*, Pracheta Sahoo *
, Yang Gao *, Anmol Agarwal*,
Kangkook Jee *, and Latifur Khan*
*Department of Computer Science
University of Texas at Dallas, Richardson, Texas 75080
Email: (gbadebo.ayoade,KhandakarAshrafi.Akbar,pracheta.sahoo
yxg122530,anmol.agarwal,Kangkook.Jee,lkhan)@utdallas.edu

Anoop Singhal†
†National Institute of Standards and Technology
Email: anoop.singhal@nist.gov

***Abstract*—Advanced persistent threats (APT) have increased in recent times as a result of the rise in interest by nation-states and sophisticated corporations to obtain high profile information. Typically, APT attacks are more challenging to detect since they leverage zero-day attacks and common benign tools. Furthermore, these attack campaigns are often prolonged to evade detection. We leverage an approach that uses a provenance graph to obtain execution traces of host nodes in order to detect anomalous behavior. By using the provenance graph, we extract features that are then used to train an online adaptive metric learning. Online metric learning is a deep learning method that learns a function to minimize the separation between similar classes and maximizes the separation between dis- similar instances. We compare our approach with baseline models and we show our method outperforms the baseline models by increasing detection accuracy on average by 11.3 % and increases True positive rate (TPR) on average by 18.3 %.**

I. Introduction

Advanced Persistent Threat (APT) [13] attacks are attacks that are usually conducted by nation state actors. These attacks target the victim's network in order to gain access to confidential information for espionage or compromise the network to destroy the victim's systems.

One example of an APT attack is the Sykipot attacks. In the Sykipot attacks, attackers targeted U.S. and U.K. organizations such as defense contractors, computer hardware manufacturers, and government departments. The attackers used spear-phishing to send emails that contained malicious attachments or links. If a user were to click on a malicious link or open a malicious attachment, then this could harm the organization's system. APT attacks are stealthy and are designed to avoid detection. Therefore, it is difficult to detect the APT attacks [10]. This is a significant challenge.

In a traditional machine learning based approach, we may train a machine learning model with class A and class B attacks, but a new attack which belongs to a novel class may suddenly appear. These novel attack classes can be termed as a zero-day attacks since the new class is not part of the training data but it may appear in the test class [16]. In this case, a traditional machine learning approach may not be able to detect the newly appeared class as a malicious attack class effectively. Due to the limitation of a traditional machine learning approach, we use a deep learning method based on Online Metric Learning (OML) [7] which learns a function to minimize the separation between similar classes (attack classes) and maximizes the separation between dissimilar instances (attack classes versus benign). Therefore, our method can recognize some of the zero-day attack more clearly from the benign instances in latent space and it performs effectively better than traditional machine learning approach. However, there is no guarantee our method will always detect all the zero-day attacks.

There has been much work regarding APT detection that attempt to address this challenge. Different methods have been proposed. For example, Milajedri et al. [17] propose the HOLMES system that gathers computer audit data and ranks the severity of the APT attack in real time. In HOLMES, an APT attack is classified based on the seven stages of the APT kill chain: 1) Initial Compromise, 2) Establish Foothold, 3) Escalate Privileges, 4) Internal Reconnaissance, 5) Move Laterally, 6) Maintain Presence, and 7) Complete Mission.

While much work has been done in the field to identify certain existing APT attacks, currently, a method to detect new APT attacks as they are being carried out does not exist. New APT attacks can only be detected after the attack had already occurred. Therefore, adversaries can still inflict significant damage when they conduct a zero-day APT attack. We propose a method to address this problem that generates an alert if a novel APT attack occurs. Our method could prevent damage from the novel APT attacks. More specifically, we train our model on a subset of the attacks, for example, shell-shock attack and test on other types of attack such as database command injection attack.

In addition, most APT attackers leverage traditional non-malicious tools to complete their attacks. For example, an attacker may exploit a bash vulnerability to open a backdoor on the victim system without installing any malicious software and can then perform lateral movement to access high target systems like the databases to steal data. By tricking the victim into running a bash script, the attacker gains access to

[1] To appear in *IEEE Conference on Communications and Network Security (CNS) 2020*, Avignon, France, Date: 29 June - 1 July, 2020

Ayoade, Gbadebo; Akbar, Khandakar; Sahoo, Pracheta; Gao, Yang; Singhal, Anoop; Jee, Kangkook; Khan, Latifur; Agarwal, Anmol. "Evolving Advanced Persistent Threat Detection Using Provenance Graph and Metric Learning." Paper presented at IEEE International Conference on Communications and Network Security (CNS 2020), Avignon, FR. June 29, 2020 - July 01, 2020.

the victim system. Detecting such attacks is challenging if the behaviour of the attack flow is not taken into consideration. In this case, a reverse-shell connection from the victim's machine to the attacker's machine takes place. A benign network flow will just involve connection to the database server from a regular network host. Our aim is to be able to detect when a non-malicious tool is used in a malicious attack flow.

Our contributions include:

- We propose and implement a system that leverages provenance graphs derived from system events for detection of APT attacks
- We propose and implement a metric learning based approach to detect novel APT attacks by learning a latent space that effectively separates benign classes from attack classes.
- We show our method OML outperforms traditional machine learning classifiers in detection of novel APT attacks by an average of 12 % in detection accuracy.

The rest of the paper is organized as follows. Section II discusses the challenges encountered in APT detection systems. Section III discusses our approach and gives an overview of the system. Section IV provides a detailed architecture of our system. Section V shows our classification method using online metric learning. Section VI presents a summary of our implementation and Section VIII shows the evaluation of our approach. Finally, Section IX provides the related work and Section XI provides a conclusion and possible future work.

II. Challenges

When designing and implementing our method, we encountered some challenges that are listed below.

- **Limited training data for novel APT attacks**: Since data for novel APT attacks is limited or non-existent, it will be difficult for us to train our machine learning model for novel APT attacks.
- **There is no signature for the zero-day APT attack**: When a novel APT attack occurs, it does not leave a specific signature that can be used to identify the attack. For novel APT attack identification, we need to analyze the victim machine's log files. Because the log files can be very large, it is challenging to filter out the attack activities from benign activities.
- **Detection in real time**: It is also challenging to detect the APT attacks in real-time. Our system needs to detect attacks quickly and alert the system users to prevent the attack from causing damage.
- **Reduce false positives**: Events that are benign could be incorrectly identified as a novel APT attack. Our goal is to accurately detect APT attacks with a low false positive rate.

III. Approach

Figure 1 shows a simplified example of a provenance graph for an example APT attack. The provenance graph is usually much larger, but for readability, we have only depicted a small portion of the graph. In this APT attack, a user goes to evilsite.com and then downloads a Trojan horse. The Trojan executes malicious commands in the background via the malicious executable file evil.exe while also providing functionality that the user is aware of. For example, a Trojan could be a malicious calculator that appears to be a normal calculator but is secretly executing malicious commands in the background as the calculator application is running. In this example, the Trojan from evilsite.com is able to successfully read the /etc/shadow file and a file containing confidential information that is named 'Top-secret'. Therefore, the Trojan is able to successfully spy on the victim and gain unauthorized access to the victim's confidential information.

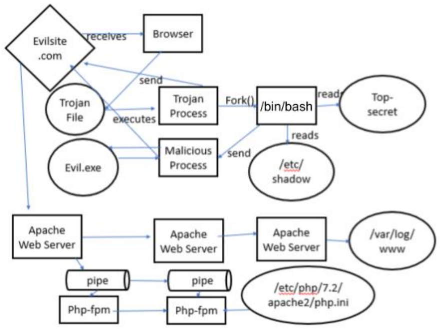


Fig. 1. Sample Camflow provenance graph data

We generate provenance graphs that look like the graph depicted in Figure 1. However, the generated provenance graphs are much larger and show all system activities and processes instead of a small snippet. For APT attack detection, we detect vulnerabilities in the provenance graph. For this, we construct a provenance labeled graph (PLG). Here, PLG is an undirected graph that is defined as $G = (V, E)$ where V is the set of vertices which include `processes`, `tasks` and, `network socket` events, $E \subset V * V$ is the set of undirected edges which include interaction between system events such as `write` and `read` events. Given a set of training examples $T = (x_i, y_i)$ where $x_i \subset X$ is a graph, and $y_i \subset Y = +1$, 1 is a target label, the graph classification problem is to induce the mapping f: $X \Rightarrow Y$. For this, after PLG extraction, we need to convert it to a feature matrix vector by using node2vec [11], GraphSAGE [12], etc. Then, we apply our novel supervised learning technique to detect APT attacks in the stream of data including novel APT attacks.

IV. Architecture

Figure 2 illustrates our approach. Our approach leverages metric learning based detection to classify unknown APT attacks in real time. We use the provenance data to visualize the activity on the machine, and then we filter this provenance data to target attack traffic. Data *provenance* is a representation of the relationships among entities (data items), activities (changes applied to the entities), and agents (people or organizations that are associated with the activities and/or

Ayoade, Gbadebo; Akbar, Khandakar; Sahoo, Pracheta; Gao, Yang; Singhal, Anoop; Jee, Kangkook; Khan, Latifur; Agarwal, Anmol. "Evolving Advanced Persistent Threat Detection Using Provenance Graph and Metric Learning." Paper presented at IEEE International Conference on Communications and Network Security (CNS 2020), Avignon, FR. June 29, 2020 - July 01, 2020.

entities) [21]. We use this provenance to gather information about all of the activities occurring on the machine that could be representative of an attack tactic.

Figure 2 illustrates the steps of our proposed solution. First, we perform simulated advanced persistent attacks on the targeted victim machine. The data from these attacks is transformed into provenance data by CamFlow [21]. Second, we convert the provenance logs generated by CamFlow to a provenance graph using the CamQuery tool [22]. Third, once we have a provenance graph, we filter out sections of the graph to generate sub graphs that contain the events that are commands executed on the system. We filter out extraneous noise that are common activities that occur on the machine regardless if the events are attacks or benign. Fourth, we build a supervised model from these training graphs to detect novel APT attacks, existing APT attacks and benign events.

For feature extraction, we convert the graph into vectors using graph embedding by leveraging node2vec. An embedding vector is learned for each unique node in the graph. The vectors for the nodes in a graph are then aggregated together using the average function for each of the instances of the attack. Our feature extraction method is further discussed in Section IV-B. We train our OML method to detect attacks based on the extracted features.

More specifically, supervised learning is utilized to incrementally learn from the data. For supervised learning, we learn accurate models by leveraging attack and benign data, which are initially gleaned from benign data, synthetic attacks and existing APT attack traces, and later from live attack detection for detecting the novel type of APT attack.

We capture the data on a machine. We capture provenance data using the tool, CamFlow [21]. The provenance data is a record of all of the activities occurring on the machine. We then generate a provenance graph using CamQuery [22]. We adopt a similar strategy to HOLMES [17], but we extend this strategy further to detect novel APT attacks.

We generate existing APT attack data by simulating the attacks on an Ubuntu Linux Virtual Machine. We simulate these attacks by performing various malicious activities on the machine such as downloading vulnerable software and running malicious programs on the machine. While we attack the machine, CamFlow [21] and CamQuery [22] capture the provenance data from the machine in the W3-PROV-JSON format and the provenance graph respectively.

A. Provenance Graph

The provenance graph is collected as a set of JSON files. Listing 1 shows an example of a node and write edge. The nodes contain the provenance type such as `fifo`, `file`, or `socket` [1]. It contains the machine id, boot id and unique node ids. The edges contain additional information such as provenance activity and entity nodes which are interacting together. In our case, we extract the interaction between the different provenance event types using CamQuery to form the nodes and edges in our graph. The Camquery extracts the node ids from the provenance graph and forms a pairwise list of the graph interaction nodes which we use as input to our feature extraction system.

```
"ABAAAAAAACAe9wIAAAAAAE7aeaI+200UAAAAAAAAAAAA="
: { "cf:id": "3",
    "prov:type": "fifo",
    "cf:boot_id": 1,
    "cf:machine_id": cf:515081690,
    "cf:version": 0,
    "cf:date": "2019:08:13T15:50:53",
    "cf:ino": 51964,
    "prov:label": "[fifo] 0" }
```

Listing 1. Sample node data for a provenance graph with FIFO type

B. Feature Extraction

Node2vec. We use Node2vec [11] to pre-process the provenance graph before classification. Node2vec is a semi-supervised algorithm for learning features from network graphs. Node2vec uses a similar approach to skip-gram by learning a vector that preserves the neighbourhood relationship of graph nodes similar to word2vec. By representing the graphs by performing a breadth-first search walk on the graph, a sequence of nodes can be generated similar to words in a document. We leverage this method to learn a vector for each node which is used for generating features for attack detection.

For our approach, we extract the node id from the provenance graph. Since the graph is a representation of how each process interacts with other processes and tasks, we model the process task as nodes and the interactions as edges. The list of edges is then passed to the node2vec algorithm to generate an embedding vector for each unique node in the graph. The embedding vector for the nodes in the graph is then combined by finding the average vector representation which is used as a feature for the attack instance.

V. Attack Detection

A. Emerging Attack Class Detection

Traditional machine learning approach is not effective at detecting novel attack classes. For example, a traditional machine learning model may be trained with instances that belong to class A and class B. However, a new attack class may emerge over time and the model may not be able to detect the new attack class effectively. In addition, in many real-world scenarios of APT attacks, instances of patterns associated with the attack type may change over time. Therefore, classifier performance is affected by the occurrences of instances from unknown or novel patterns.

For our novel class detection, we leverage online metric learning or distance based learning where malicious instance points can be well separated from benign class instances so that when novel attack classes emerge, we can detect it effectively. We provide more discussion in the next section.

B. Online Metric Learning

Online Adaptive Metric Learning (OAML) [9] [5] [4] is based on a deep learning architecture that transforms an instance feature from an original feature space to a latent feature space. By transforming to a latent feature space,

Ayoade, Gbadebo; Akbar, Khandakar; Sahoo, Pracheta; Gao, Yang; Singhal, Anoop; Jee, Kangkook; Khan, Latifur; Agarwal, Anmol. "Evolving Advanced Persistent Threat Detection Using Provenance Graph and Metric Learning." Paper presented at IEEE International Conference on Communications and Network Security (CNS 2020), Avignon, FR. June 29, 2020 - July 01, 2020.

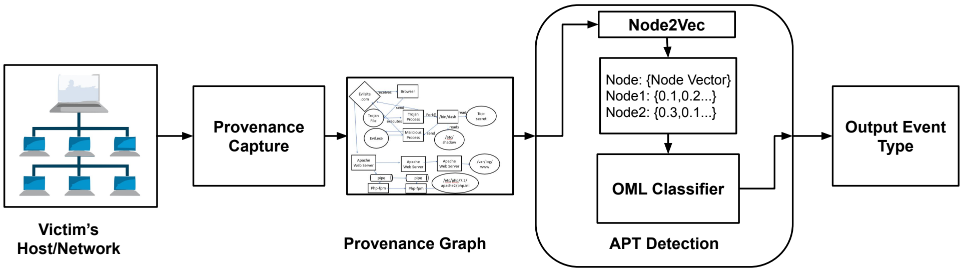


Fig. 2. Architecture

the metric distance between dissimilar instances is increased and distance between similar classes is reduced. The work leverages methods which use *pairwise* and *triplet* constraints.

Our OAML method learns a non-linear similarity metric unlike others which uses a pre-selected linear metric (e.g., Mahalanobis distance [26]). Our OAML method overcomes bias to a specific dataset by using an adaptive learning method. Our OAML leverages neural networks where the hidden layer output is passed to an independent metric-embedding layer (MEL). The MELs then generate an n-dimensional embedding vector as output in different latent space.

1) Problem Setting: Let $S = \{(x_t, x_t^+, x_t^-)\}_{t=1}^T$ be a sequence of triplet constraints sampled from the data, where $\{x_t, x_t^+, x_t^-\} \in \mathcal{R}^d$, and x_t (anchor) is similar to x_t^+ (positive) but dissimilar to x_t^- (negative). The goal of online adaptive metric learning is to learn a model $F : \mathcal{R}^d \mapsto \mathcal{R}^{d'}$ such that $||F(x_t) - F(x_t^+)||_2 \ll ||F(x_t) - F(x_t^-)||_2$. Given these parameters, the objective is to learn a metric model with adaptive complexity while satisfying the constraints. The complexity of F must be adaptive so that its hypothesis space is automatically modified.

2) Overview: Consider a neural network with L hidden layers, where the input layer and the hidden layer are connected to an independent MEL. Each embedding layer learns a latent space where similar instances are clustered and dissimilar instances are separated.

Figure 3 illustrates our Artificial Neural Network (ANN) . Let $E_\ell \in \{E_0, E_1, E_2, \ldots, E_L\}$ denote the ℓ^{th} metric model in OAML (i.e., the network branch from the input layer to the ℓ^{th} MEL). The simplest OAML model E_0 represents a linear transformation from the input feature space to the metric embedding space. A weight $\alpha^{(\ell)} \in [0, 1]$ is assigned to E_ℓ, measuring its importance in OAML.

For a triplet constraint (x_t, x_t^+, x_t^-) that arrives at time t, its metric embedding $f^{(\ell)}(x_t^*)$ generated by E_ℓ is

$$f^{(\ell)}(x_t^*) = h^{(\ell)}\Theta^{(\ell)} \quad (1)$$

where $h^{(\ell)} = \sigma(W^{(\ell)}h^{(\ell-1)})$, with $\ell \geq 1$, $\ell \in \mathbb{N}$, and $h^{(0)} = x_t^*$. Here x_t^* denotes any anchor (x_t), positive (x_t^+), or negative (x_t^-) instance, and $h^{(\ell)}$ represents the activation of the ℓ^{th} hidden layer. Learned metric embedding $f^{(\ell)}(x_t^*)$ is limited to a unit sphere (i.e., $||f^{(\ell)}(x_t^*)||_2 = 1$) to reduce the search space and accelerate training.

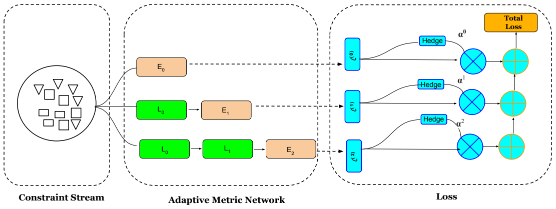


Fig. 3. OAML network structure consist of L_i linear layer and Embedding layers E_i layer.

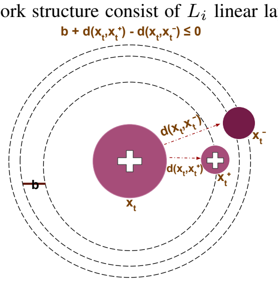


Fig. 4. Data instance before applying Online metric learning

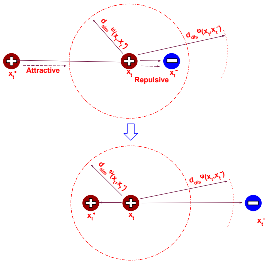


Fig. 5. Data instance after projection using OML

During the training phase, for every arriving triplet (x_t, x_t^+, x_t^-), we first retrieve the metric embedding $f^{(\ell)}(x_t^*)$ from the ℓ^{th} metric model using Eq. 1. A local loss $\mathcal{L}^{(\ell)}$ for E_ℓ is evaluated by calculating the similarity and dissimilarity errors based on $f^{(\ell)}(x_t^*)$. Thus, the overall loss introduced by this triplet is given by

$$\mathcal{L}_{overall}(x_t, x_t^+, x_t^-) = \sum_{\ell=0}^{L} \alpha^{(\ell)} \cdot \mathcal{L}^{(\ell)}(x_t, x_t^+, x_t^-) \quad (2)$$

Parameters $\Theta^{(\ell)}$, $\alpha^{(\ell)}$, and $W^{(\ell)}$ are learned during the online learning phase. The final optimization problem to solve in OAML at time t is therefore:

$$\begin{aligned} &\underset{\Theta^{(\ell)}, W^{(\ell)}, \alpha^{(\ell)}}{\text{minimize}} && \mathcal{L}_{overall} \\ &\text{subject to} && \|f^{(\ell)}(x_t^*)\|_2 = 1, \forall \ell = 0, \ldots, L. \end{aligned} \quad (3)$$

We evaluate the similarity and dissimilarity errors using an *adaptive-bound triplet loss* (ABTL) constraint [9] to estimate $\mathcal{L}^{(\ell)}$ and update parameters $\Theta^{(\ell)}$, $W^{(\ell)}$ and $\alpha^{(\ell)}$.

3) Why OML works: A typical machine learning algorithm like k-NN will misclassify the instances shown in Figure 4, since x_t^+ is closer to x_t^- and further from x_t. In our case, most APT attacks use non-malicious software to complete their attack activities making the attack events and traffic look non-malicious. To overcome this challenge, we use ABTL.

Figure 5 illustrates the main idea of ABTL. The objective is to have the distance $D_{update}^{(l)}(x_t, x_t^+)$ of two similar instances x_t and x_t^+ to be less than or equal to a similarity threshold $d_{sim}^{(l)}(x_t, x_t^+)$ so that the attractive loss $\mathcal{L}_{attr}^{(l)}(x_t, x_t^+)$ drops to zero; on the other hand, for two dissimilar instances x_t and x_t^-, we desire their distance $D_{update}^{(l)}(x_t, x_t^-)$ to be greater than or equal to a dissimilarity threshold $d_{dis}^{(l)}(x_t, x_t^-)$, thereby reducing the repulsive loss $\mathcal{L}_{rep}^{(l)}(x_t, x_t^-)$ to zero.

4) Adaptive-Bound Triplet Loss: Here $y_t \in \{+1, -1\}$ denotes whether x_t is similar $(+1)$ or dissimilar (-1) to x_t' and $b \in \mathcal{R}$ is a user-specified fixed margin. While triplet loss simultaneously learns both similarity and dissimilarity relations, the pairwise loss can only focus on one of the relations at a time, which leads to a poor metric quality. In addition, triplet loss requires a proper margin be specified. In addition, the selected margin is highly dependent on the data and it requires extensive domain knowledge. Our aim is to automatically learn the margin for our triplet-loss constraint irrespective of the available data.

With $\tau \in (0, \frac{2}{3})$, by optimizing the proposed adaptive-bound triplet loss, different classes are separated in the metric embedding space. Let $D(c_1, c_2)$ denote the minimal distance between classes c_1 and c_2, i.e., the distance between two closest instances from c_1 and c_2 respectively. Consider an arbitrary quadruple $(x_1, x_2, x_3, x_4) \in \mathcal{Q}$ where $\{x_1, x_2\} \in c_1$, $\{x_3, x_4\} \in c_2$, and $\mathcal{Q}$ is the set of all possible quadruples generated from class c_1 and c_2. Suppose (x_2, x_3) is the closest dissimilar pair among all possible dissimilar pairs that can be extracted from (x_1, x_2, x_3, x_4). We first prove that the lower bound of $D(c_1, c_2)$ is given by $\min\limits_{(x_1,x_2,x_3,x_4) \in \mathcal{Q}} D^{(l)}(x_1, x_4) - D^{(l)}(x_1, x_2) - D^{(l)}(x_3, x_4)$.

$$\begin{aligned} D^{(l)}(x_1, x_4) &\leq D^{(l)}(x_1, x_2) + D^{(l)}(x_2, x_4) \\ &\leq D^{(l)}(x_1, x_2) + D^{(l)}(x_2, x_3) + D^{(l)}(x_3, x_4) \end{aligned} \quad (4)$$

$$\begin{aligned} D(c_1, c_2) &= \min_{(x_1,x_2,x_3,x_4) \in \mathcal{Q}} D^{(l)}(x_2, x_3) \\ &\geq \min_{(x_1,x_2,x_3,x_4) \in \mathcal{Q}} D^{(l)}(x_1, x_4) - D^{(l)}(x_1, x_2) \\ &-D^{(l)}(x_3, x_4) \end{aligned} \quad (5)$$

By optimizing the adaptive-bound triplet loss, the following constraints are satisfied.

$$\begin{cases} D^{(l)}(x_1, x_2) \leq d_{sim}^{(l)}(x_1, x_2) \leq \mathcal{T}_{sim}^{(l)} \\ D^{(l)}(x_3, x_4) \leq d_{sim}^{(l)}(x_3, x_4) \leq \mathcal{T}_{sim}^{(l)} \\ D^{(l)}(x_1, x_4) \geq d_{dis}^{(l)}(x_1, x_4) \geq \mathcal{T}_{dis}^{(l)} \end{cases} \quad (6)$$

$$\begin{aligned} D(c_1, c_2) &\geq \min_{(x_1,x_2,x_3,x_4) \in \mathcal{Q}} D^{(l)}(x_1, x_4) - D^{(l)}(x_1, x_2) \\ &-D^{(l)}(x_3, x_4) \\ &\geq \mathcal{T}_{dis}^{(l)} - 2\mathcal{T}_{sim}^{(l)} \\ &= 2 - \tau - 2\tau \\ &= 2 - 3\tau \end{aligned} \quad (7)$$

if $\tau \in (0, \frac{2}{3})$, we have $3\tau < 2$. Therefore,

$$D(c_1, c_2) \geq 2 - 3\tau > 0 \quad (8)$$

Equation 8 indicates that the minimal distance between class c_1 and c_2 is always positive so that these two classes are separated.

Note that our whole proof is solely based on the triangle inequality, which is correct as long as the triangle inequality holds. As L_2-norm is utilized to measure the distance, the metric space learned by our framework is indeed a normed vector space. The triangle inequality is a natural property in this space.

Moreover, our framework does not require the learned metric space to be convex, i.e., for any two distinct instances x and y in the metric space, there exists a third instance z in this space lying between x and y ($d(x,z)+d(z,y) = d(x,y)$). Whether or not the equality strictly holds does not affect the correctness of our proof, since it only considers the upper-bound of a distance. In our case, even the non-convex metric space is a valid solution, as the triangle inequality still holds in this space.

VI. Implementation

We developed an implementation of our system for the 64-bit Ubuntu Linux operating system with 128 GB space of RAM. Our system consists of the data generation and the machine learning detection module. For the data generation, we used bash scripts which consist of 100 lines of code. We leveraged the CamFlow tool [21] installed on VirtualBox using Vagrant. We modified the CamQuery module [22] to extract the ID of the provenance graph nodes and to generate edges of process interactions.

For the machine learning detection module, we leveraged scikit-learn for our baseline evaluation and the OAML components were implemented with approximately 500 lines of Python code using the PyTorch library.

Model Parameters. For our experiments, Support Vector Machine (SVM) uses Radial Basis Function (RBF) kernel with Cost = 1.3×10^5 and gamma is set to 1.9×10^{-6}. OAML uses a Rectified Linear activation Unit (ReLU) network with embedding $n = 200$, number of hidden layers $L = 5$, $k = 1$, learning rate = 0.3, learning rate decay = 1×10^{-4}, and uses ADAM (A method for stochastic optimizer) optimizer.

VII. Datasets

A. Attack Generation & Data Collection

Table I show the contents of our dataset. The dataset consists of different APT activities such as exfiltration of data, illegal login and access, opening of a reverse shell for command and control access, and illegal network scanning using **nmap** for the discovery of services on victim's network. All the classes were used for both benign and malicious scenario generation. For example, command line injection attacks were considered malicious when some third party injected some commands and executed them. Benign scenarios were mimicked with normal command execution flow which consists of usual executed commands. Table II shows properties of the provenance graph based on the trace of a single execution of an attack instance. A sample provenance graph contains an average of 10 332 edges for data ex-filtration attack events labeled as class 1, while the average in-degree is 48 and the average out-degree is 153 for the same class.

For our dataset, we collect the provenance graph which contains information such as provenance type and task type as discussed in section IV-A. Benign instances include web browsing activities on benign websites, normal login activity with Secure Shell (SSH) and benign connection from a client machine to a database server. We collected 100 traces for each attack type. Our dataset consists of activities derived from previously known APT attack campaign steps. We collect multiple traces of attack sequences to gather diversified training data and test on single instance of the attack sequence. We generated 7 benign cases and 7 attack cases since each attack type has a corresponding benign activity variation.

B. Attack Generation

Table I shows that the Data Leakage Attack & Remote Webservice Penetration (Shellshock) and Password Cracking attack are deployed through mimicking real-world scenarios. A server end user, a client end user, and an attacker are required to mimic the attack. In the data leakage attack scenario, a server consists of a database and owns it and has given access to its client to access data from the database through queries. In the server end, postgresql database is used for such service. If a database is dumped from the server end into any sql file or any other file format, a client can reconstruct the database or restore the database by having that file from the server end. In the attack scenario, a third user who is essentially the attacker makes the client download some malicious program (e.g., spear phishing e-mail (from outside) to open a backdoor or sending the software to the client end by any other means) and run it in the client's computer. With the execution of that software, the attacker gets access to the client's system (control over client's computer) and can now communicate with the server. The attacker can now make queries to the server and can even restore the database in its own end but it will make the behavior seem like a client as it will be doing these tasks from the client's system. In a nutshell, the attacker makes the victim end download some executable and execute it to open the backdoor for itself to the server end so that it can leak the data to the client end while having full access to the server as a client (a client acts as a privileged domain to the attacker). The attacker can leak the data through external storage or network transfer after the attack has been deployed from the client's computer. Benign scenarios for this case were created with normal operational and communication flow between the server and the client. These operations include normal database queries made by the client to the server, access to the database file and retrieving of the database by the client itself.

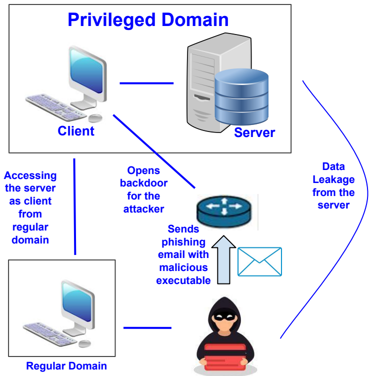

Privileged Domain
Client
Server
Accessing the server as client from regular domain
Opens backdoor for the attacker
Sends phishing email with malicious executable
Data Leakage from the server
Regular Domain

Fig. 6. Data Leakage Attack

The Remote Webservice Penetration (Shellshock) and Password Cracking attack is deployed when the client's system is somewhat hacked or deliberately used to attack the server end. In this scenario, an attacker exploits a remote shellshock vulnerability by sending crafted input to the CGI (Common Gateway Interface) service implemented using bash. Thus the attacker leverages the shellshock vulnerability as a backdoor to run malicious commands such as executing a password cracking tool.

C. Data Collection

1) Data Collection for Data Leakage Attack: Figure 6 illustrates the steps required to complete this attack. For the data leakage attack, data was captured in both the server and client end. For the attack scenario, the attacker sends an email with some malicious executable attached with it to the client's mail. The client downloads the attachment, executes it, and thus opens the backdoor for the attacker without any knowledge of it. The attacker now mimics the behavior of the client but from its own end and performs all the database queries and restoration. When the client executes the backdoor program, it goes straight to a vulnerable state

TABLE I
SUMMARY OF ATTACK WORKLOAD

Class	Description	Software
1	Data exfiltration	scp
2	Illegal network scanning	ping
3	Illegal network mapping	nmap
4	Reverse shell for Command and Control	nc
5	Data Leakage Attack	Deployed Through Mimicking Real-World Scenario
6	Remote Webservice Penetration (Shellshock) and Password Cracking	Deployed Through Mimicking Real-World Scenario
7	Command Line injection Attack	Deployed Through Mimicking Real-World Scenario

Attacks emulated from MITRE APT collection.

TABLE II
PROVENANCE GRAPH PROPERTIES

Class	Avg # of out-deg	Avg # of in-deg	Avg # of edges
1	153	48	10 337
2	545	13	9796
3	325	5	9781
4	511	96	10 236
5	561	922	9783
6	2,084	1,437	10 638
7	300	52	10 106

and from that time Camflow provenance data is collected both at the server and client end. The benign set of data corresponds to when the server and client are communicating regularly. No other connection is established in the meantime. The server receives database requests from the client in a regular manner in benign scenario data collection.

2) Data Collection for Password Cracking Attack: We carried out a password cracking attack using the John the ripper tool which implements a dictionary password attack. For the password cracking attack, the normal behavior or benign scenario simply involves a client connecting with the server using a curl command with no malicious payload. Benign data is collected while these normal operations take place using a Camflow provenance graph. For the attack scenario, the attacker exploits the shellshock vulnerability and then downloads a password cracker and executes a shell script for the backdoor creation. The attacker then executes commands for password collection and cracking from the victim's end. While these steps take place, data was collected at the victim's end using a Camflow provenance graph.

3) Data Collection for Command Line Injection Attack: For the command line injection attack, we only collected data for the attack scenario. The victim host (e.g., say embedded / IoT device that runs Linux) runs Mediaplayer (Kodi client), and it exports a remote control Application Program Interface (API) as a web service. One of its input sanitizations has an error that fails to filter invalid input from the outside, in turn allowing attackers to inject arbitrary commands blended in one of its requests. This attack is inspired by the Jeep-Cherokee attack case where the attacker from remote gains control over the vehicle. In parallel to the above steps, the data collection was done at the victim host using a Camflow provenance graph.

TABLE III
RESULTS BASED ON NOVEL APT ATTACK DETECTION

Metric	Classes #	OML	KNN	SVM_RBF	Decision_Tree
Accuracy	1	**54.76**	48.02	40.03	47.95
	2	**64.99**	56.67	40.03	66.64
	3	**73.18**	72.66	50.2	60.7
	4	**75.96**	72.85	61.36	66.91
	5	**86.06**	79.52	68.03	73.51
	6	**92.07**	86.2	71.0	80.52
	7	**98.08**	92.87	77.48	87.45
	8	**98.**	96.37	91.88	93.66
	9	99	99	91.88	99
Metric	**Classes #**	**OML**	**KNN**	**SVM_RBF**	**Decision_Tree**
F2	1	**28.92**	16.12	0	15.99
	2	**47.13**	32.44	0	49.94
	3	**60.72**	59.86	22.02	39.67
	4	**65.13**	60.18	42.62	50.38
	5	**80.5**	70.69	53.91	61.25
	6	**89.14**	80.7	58.77	72.2
	7	**97.43**	90.25	68.98	82.53
	8	**97.46**	95.09	90.09	91.36
	9	99	1	90.09	99

TABLE IV
RESULTS BASED ON NOVEL APT ATTACK DETECTION

Metric	Classes #	OML	KNN	SVM_RBF	Decision_Tree
FPR	1	0	0	0	0
	2	0	0	0	0
	3	0	0	0.0231	0
	4	0	0	0.0281	0
	5	0	0	2.81	0
	6	0	0	2.81	0
	7	0	0	2.81	0
	8	0	0	2.81	0
	9	0	0	2.81	0
Metric	**Classes #**	**OML**	**KNN**	**SVM_RBF**	**Decision_Tree**
TPR	1	**24.56**	13.33	0	13.22
	2	**41.63**	27.75	0	44.38
	3	**55.29**	54.41	18.5	34.47
	4	**59.91**	54.74	37.44	44.82
	5	**76.76**	65.86	48.57	55.84
	6	**86.78**	76.98	53.52	67.51
	7	**96.81**	88.11	64.32	79.07
	8	**97.96**	93.94	88.33	89.43
	9	99	1	88.33	99

VIII. EVALUATION

A. Results

Table III and IV show the results of our experiment. For our experiment, we trained incrementally on the attack classes and tested on all the remaining classes which includes the benign class instances. Please note that the training data and the test data consists of the benign data instances. With this approach, we can determine if our algorithm can detect unseen novel attack classes. First, we train on all the benign classes and a single APT attack class and test on all the benign and APT attack classes. Second, we train on all benign

TABLE V
EXECUTION TRAINING AND TESTING TIME FOR OML

No of Training Instances	Train time(s)	Test time(s)
700	55.8	1.94
800	55.9	1.95
900	55.4	1.96
1000	55.5	1.96
1100	54.7	1.91
1200	55.9	1.97
1300	58.9	1.98
1400	59.5	2.00
1500	59.6	2.08

classes and two APT attack classes and tested on all attack classes. Third, we train on all benign classes and three APT attack classes and tested on all attack classes. Lastly, we train on all benign classes and all the APT attack classes and tested on all attack classes. We measured our performance by using the following metrics: **Accuracy**, **TPR**, **FPR**, and **F2**.

In the experiments, we measured the true positive rate (*tpr*), where true positive represents the number of correctly classified seen and novel APT attacks classes; false positive rate (*fpr*), where false positive represents the number of incorrectly classified seen and novel APT attacks; and F_2 score of the classifier, where the F_2 score is interpreted as the weighted average of the precision and recall. The F2 score ranges between the values 100 and 0 where 100 is the best value and 0 is the worst value.

The results show our approach performs better than traditional machine learning based classifiers such as k-NN, SVM and Decision tree. The accuracy of detection of novel attacks with our approach is 86 % compared to 79 % for k-NN, 68 % for SVM and 73 % for Decision Tree when we train on only five attack classes. Similarly, the TPR is 76 % for OML compared to 65 % for k-NN and 55 % for Decision Tree when we train on only five attack classes. The $F2$ is 80.5 % for OML compared to 70.69 % for k-NN and 61 % for Decision Tree when we train on only five attack classes. The $accuracy$ increases to 98 % for OML and 96 % for k-NN, 91 % for SVM and 93 % for Decision Tree when we train on 7 attack classes.

Our approach improves classification performance on average by **6.8 %** for accuracy, **10.19 %** for TPR and **11.4 %** for $F2$ when compared with k-NN method, while the performance improves by **17 %** for accuracy, **28 %** for TPR and **26 %** for $F2$ when compared with SVM. Likewise, our performance improves by **10 %** for accuracy, **17 %** for TPR and **16 %** for $F2$ when we compared our method with Decision Tree.

Our approach detected novel APT attacks with higher accuracy than k-NN, SVM or Decision Tree even with limited training on a subset of the APT attack classes. This is possible because OML can learn to minimize the distance in feature space for similar instances and maximize the distance for dissimilar instances as shown in Theorem V-B4.

B. Execution time for OML

Table V shows the summary of the execution time required to train our OML approach and then perform inference on the test data. The execution time for training the OML algorithm is approximately 60 seconds while the testing or inference execution time is approximately 2 seconds. As discussed in subsection V-B, OML learns a latent embedding space vector to satisfy a constraints. As a result of this algorithm, OML training time is higher than at inference. As we can see from the execution timing information, the testing time is fast as a result of just using the learned embedding vector to classify the test data. In addition, we only show the number of instances used for the training since the number of testing instances is always constant as discussed in section VIII.

IX. RELATED WORK

Our work specifically focuses on APT attack detection. An earlier approach was proposed by [23]. In their work they coined an automated technique to generate attack graphs using symbolic model checking algorithms. Later on [24] proposed a robust , flexible graph based approach for network vulnerability analysis which allows attacks from both outside and inside the networks. This method suffers from the scalability issue when state increase occurs. [3] proposed scalable attack graphs which do not require the idea of backtracking from the attacker side relying on the idea of monotonicity. [18] proposed an idea based on Decision Tree to prevent APT attacks. Recent work on APT attack detection include Milajerdi et al. [17] called HOLMES. HOLMES uses a set of manually generated rules to describe different APT information flows from an attack provenance graph. Our approach uses a deep learning based approach to learn the attack patterns without the need for manual generation of rules. Ghafir et al. [10] uses machine learning based correlation analysis MLAPT, but the approach does not focus on detection of novel attacks.

There are additional works that propose various methods for APT detection. For example, [6] proposes a novel deep learning stack for APT detection. In this method, they propose using deep learning for outlier detection to detect APT attacks and novel APT attacks. However, this method has not been implemented in practice. Cho et al. [8] propose a method for APT detection based on unusual or unknown domains that a malicious user could visit while conducting an APT attack. In [14] authors have proposed STREAMSPOT, a clustering based anomaly detection method in heterogeneous graph. Through this method, anomalous graphs, that are prominently different from others are identified. Tian et al. [25] proposed a deep learning representation for graph clustering.

Various methods for detecting attacks have been proposed. [15] uses host and network data for anomaly detection. Methods such as [20], [2] detect malware in evolving data streams. [19] uses a deep auto-encoder to learn features for novel class detection. Our work differs from these works by using provenance graphs, which take into consideration the information flow for the attack events.

Ayoade, Gbadebo; Akbar, Khandakar; Sahoo, Pracheta; Gao, Yang; Singhal, Anoop; Jee, Kangkook; Khan, Latifur; Agarwal, Anmol. "Evolving Advanced Persistent Threat Detection Using Provenance Graph and Metric Learning." Paper presented at IEEE International Conference on Communications and Network Security (CNS 2020), Avignon, FR. June 29, 2020 - July 01, 2020.

X. Limitation and Future Work

In this work, we have not focused on multi-stage attack detection, however, our framework is able to detect a single stage in a APT attack as shown in the experiments. However, our method may not always detect all the zero-day attacks. In future work, we plan to address the challenge of multi-stage attack detection. Data collection is still challenging in cyber-attack space as the amount of standardized training data is limited. In future work, we opt to collect more data to train our machine learning models. In addition, we will perform experiments with real life attack data. Our work currently focuses on attacks based on Linux operating system, we plan to address attacks in other operating systems e.g. Windows operating systems platform.

XI. Conclusion

Detecting APT attacks is a challenging task based on the approach deployed by malicious actors. In this work, we focus on a machine learning based approach to detect APT attacks. We leverage provenance graphs for the collection of event data from host systems. We apply OML: a novel machine learning technique for detecting APT attacks. Our results show our approach has a higher detection accuracy compared to traditional machine learning techniques. In our future work, we will explore more novel detection machine learning methods to detect novel APT attacks.

Acknowledgments

The research reported herein was supported in part by ONR award N00014-17-1-2995; NSA award H98230-15- 1-0271; AFOSR award FA9550-14-1-0173; NSF FAIN awards DGE-1931800, OAC-1828467, and DGE-1723602; NSF awards DMS-1737978 and MRI-1828467; an IBM faculty award (Research); and an HP grant. Any opinions, recommendations, or conclusions expressed are those of the authors and not necessarily of the aforementioned supporters.

Disclaimer. This paper is not subject to copyright in the United States. Commercial products are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

References

[1] "Camflow - vertices supported by camflow," https://github.com/CamFlow/camflow-dev/blob/master/docs/VERTICES.md.

[2] T. Al-Khateeb, M. M. Masud, L. Khan, C. Aggarwal, J. Han, and B. Thuraisingham, "Stream classification with recurring and novel class detection using class-based ensemble," in *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ser. ICDM '12, USA, 2012, p. 31–40.

[3] P. Ammann, D. Wijesekera, and S. Kaushik, "Scalable, graph-based network vulnerability analysis," in *Proceedings of the 9th ACM Conference on Computer and Communications Security*. ACM, 2002, pp. 217–224.

[4] F. Araujo, G. Ayoade, K. Al-Naami, Y. Gao, K. W. Hamlen, and L. Khan, "Improving intrusion detectors by crook-sourcing," in *Proceedings of the 35th Annual Computer Security Applications Conference*, San Juan, Puerto Rico, December 2019, pp. 245–256.

[5] G. Ayoade, F. Araujo, K. Al-Naami, A. M. Mustafa, Y. Gao, K. W. Hamlen, and L. Khan, "Automating cyberdeception evaluation with deep learning," in *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Grand Wailea, Maui, January 2020.

[6] T. Bodström, T.; Hämäläinen, "A novel deep learning stack for apt detection." *Appl. Sci.*, vol. 9, no. 1055, 2019.

[7] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

[8] D. X. Cho and H. H. Namb, "A method of monitoring and detecting apt attacks based on unknown domains," *Procedia Computer Science*, vol. 150, pp. 316–323, 2019.

[9] Y. Gao, Y.-F. Li, S. Chandra, L. Khan, and B. Thuraisingham, "Towards self-adaptive metric learning on the fly," in *The World Wide Web Conference*. ACM, 2019, pp. 503–513.

[10] I. Ghafir, M. Hammoudeh, V. Prenosil, L. Han, R. Hegarty, K. Rabie, and F. J. Aparicio-Navarro, "Detection of advanced persistent threat using machine-learning correlation analysis," *Future Generation Computer Systems*, vol. 89, pp. 349–359, 2018.

[11] A. Grover and J. Leskovec, "Node2vec: Scalable feature learning for networks," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 855–864. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939754

[12] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 1024–1034. [Online]. Available: http://papers.nips.cc/paper/6703-inductive-representation-learning-on-large-graphs.pdf

[13] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," *Leading Issues in Information Warfare & Security Research*, vol. 1, p. 80, 2011.

[14] E. Manzoor, S. M. Milajerdi, and L. Akoglu, "Fast memory-efficient anomaly detection in streaming heterogeneous graphs," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1035–1044.

[15] M. Masud, L. Khan, and B. Thuraisingham, *Data Mining Tools for Malware Detection*. CRC Press, 2011.

[16] M. M. Masud, T. M. Al-Khateeb, K. W. Hamlen, J. Gao, L. Khan, J. Han, and B. Thuraisingham, "Cloud-based malware detection for evolving data streams," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 3, Oct. 2008.

[17] S. M. Milajerdi, R. Gjomemo, B. Eshete, R. Sekar, and V. Venkatakrishnan, "Holmes: Real-time apt detection through correlation of suspicious information flows," in *2019 IEEE Symposium on Security and Privacy*, vol. 1. IEEE, 2019, pp. 447–462.

[18] D. Moon, H. Im, I. Kim, and J. H. Park, "Dtb-ids: an intrusion detection system based on decision tree using behavior analysis for preventing apt attacks," *The Journal of supercomputing*, vol. 73, no. 7, pp. 2881–2895, 2017.

[19] A. M. Mustafa, G. Ayoade, K. Al-Naami, L. Khan, K. W. Hamlen, B. Thuraisingham, and F. Araujo, "Unsupervised deep embedding for novel class detection over data stream," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1830–1839.

[20] P. Parveen, J. Evans, B. Thuraisingham, K. W. Hamlen, and L. Khan, "Insider threat detection using stream mining and graph mining," in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 2011, pp. 1102–1110.

[21] T. Pasquier, X. Han, M. Goldstein, T. Moyer, D. Eyers, M. Seltzer, and J. Bacon, "Practical whole-system provenance capture," in *Proceedings of the 2017 Symposium on Cloud Computing*, ser. SoCC '17. New York, NY, USA: ACM, 2017, pp. 405–418. [Online]. Available: http://doi.acm.org/10.1145/3127479.3129249

[22] T. Pasquier, X. Han, T. Moyer, A. Bates, O. Hermant, D. Eyers, J. Bacon, and M. Seltzer, "Runtime analysis of whole-system provenance," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18. New York, NY, USA: ACM, 2018, pp. 1601–1616. [Online]. Available: http://doi.acm.org/10.1145/3243734.3243776

[23] O. Sheyner, J. Haines, S. Jha, R. Lippmann, and J. M. Wing, "Automated generation and analysis of attack graphs," in *Proceedings 2002 IEEE Symposium on Security and Privacy*. IEEE, 2002, pp. 273–284.

[24] L. P. Swiler and C. Phillips, "A graph-based system for network-vulnerability analysis," Sandia National Labs., Albuquerque, NM (United States), Tech. Rep., 1998.

[25] F. Tian, B. Gao, Q. Cui, E. Chen, and T.-Y. Liu, "Learning deep representations for graph clustering," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[26] S. Xiang, F. Nie, and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.

RECOMMENDED OPTIONS FOR IMPROVING THE FUNCTIONAL RECOVERY OF THE BUILT ENVIRONMENT

S. Sattar(1), M. Mahoney(2), R. Kersting(3), J. Heintz(4), K. Johnson(5), L. Arendt(6), C. Davis(7), Pataya Scott(8), Leslie Abrahams(9)

(1) Research Structural Engineer, National Institute of Standards and Technology, siamak.sattar@nist.gov
(2) Senior Geophysicist, Federal Emergency Management Agency, mike.mahoney@fema.dhs.gov
(3) Associate Principal, Buehler, rkersting@buehlerengineering.com
(4) Executive Director, Applied Technology Council, jheintz@atcouncil.org
(5) Social Scientist, National Institute of Standards and Technology, katherine.johnson@nist.gov
(6) Professor, Saint Norbert College, lucy.arendt@snc.edu
(7) Consultant, cadavisengr@yahoo.com
(8) Civil Engineer, Federal Emergency Management Agency, pataya.scott@fema.dhs.gov.
(9) Research Staff Member, Institute for Defense Analyses, Science and Technology Policy Institute, labraham@ida.org

Abstract

During a strong earthquake, commercial and residential buildings designed to meet current building codes and standards may sustain damage that significantly hinders the restoration of building functionality. Similarly, lifeline infrastructure systems can be damaged and loose ability to provide critical services. The impacted buildings and lifeline infrastructure systems and their associated consequences such as dislocation of people, disruption of key services, and lack of access to jobs and schools, pose a significant impediment for communities during recovery. In the wake of recent disasters across the globe, there is mounting evidence that the public finds these kinds of disruptions unacceptable. Buildings and lifeline infrastructure systems can be designed for higher performance so that they are able to serve their function or regain functionality with acceptable interruption after an earthquake. This will require shifts in design philosophy from safety-based objectives to recovery-based objectives across multiple aspects of current practice. The 2018 Congressional reauthorization of the National Earthquake Hazards Reduction Program (NEHRP), P.L. 115-307, requires two Federal agencies, the National Institute of Standards and Technology (NIST) and the Federal Emergency Management Agency (FEMA), to work with experts across the U.S. to address this issue of improving post-earthquake functional recovery. FEMA and NIST convened a committee of experts to develop the report to the U.S. Congress to identify and assess options for functional recovery and post-earthquake re-occupancy. As part of this effort, stakeholder workshops were held in February 2020 in five U.S. cities to gather public feedback to inform the final report to Congress. This paper summarizes the main findings of this effort, including a list of the key recommendations identified for improving the functional recovery of buildings and infrastructure. The report to the U.S. Congress serves as a starting point for improving timeframes for re-occupancy and functional recovery of the built environment and critical infrastructure after earthquakes through the development and adoption of functional recovery concepts, codes and standards, policies, and practice.

Keywords: Functional recovery, post-earthquake re-occupancy, NEHRP.

Sattar, Siamak; Johnson, Katherine. "Recommended Options for Improving the Functional Recovery of the Built Environment and Critical Infrastructure." Paper presented at 17 World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

1. Introduction

Earthquakes can affect communities through loss of life, injury, property damage, displacement of residents and businesses, and long-lasting economic and social impacts. The United States Geological Survey (USGS) estimates nearly half of Americans are at risk from potentially damaging earthquakes[1]. Despite decades of improvements in the seismic safety of the built environment, the economic and social systems of communities throughout the U.S. remain at risk of large scale, long-term disruption. The U.S. has not experienced a major damaging earthquake since 1994. However, the Federal Emergency Management Agency (FEMA) estimates the annualized cost of damage to U.S. building stock from earthquakes to be $6.1B per year [2]. A 2008 USGS exercise using a 7.8M earthquake in Southern California as a case study estimated 2,000 deaths, 50,000 injuries, and $200 billion in direct costs, in addition to staggering, destabilizing impacts to long-term community function. Depending on the size of the disaster, communities can also face significant and costly long-term consequences, including interruption of basic services (shelter, food, water, sanitation), loss of jobs and businesses, voluntary and forced relocation of residents, psychological trauma, and loss of important physical, cultural, and social assets [3].

The Federal Government has recognized the need to improve the state of practice in design and retrofit of multiple components of the built environment such that buildings and lifeline infrastructure systems can restore their function with minimum disruption in the services that they provide. Buildings and lifeline infrastructure support and enable society's continued economic, psychological, and social health, all of which may be severely interrupted or completely disrupted depending upon the magnitude of shaking and level of damage experienced. By strengthening the ability of the built environment to withstand earthquake effects, we can return community members to their homes, business, and normal activities more quickly. To move toward this desired performance state, the National Institute of Standards and Technology (NIST) and the Federal Emergency Management Agency (FEMA), as part of the December 2018 reauthorization of the National Earthquake Hazards Reduction Act, were charged to convene a committee of experts *"to assess and recommend options for improving the built environment and critical infrastructure to reflect performance goals stated in terms of post-earthquake re-occupancy and functional recovery time.*", P.L. 115-307, This paper summarizes the effort to fulfill this congressional mandate and presents some key findings on the recommended options for improving the recovery time.

2. Problem Statement

In the United States, building codes are the primary mechanism by which State and local jurisdictions manage earthquake risk for buildings. Most often, these building codes are adopted with the target of saving lives and reducing injuries, rather than preserving the structure's ability to be operational, or even recoverable, after an earthquake event. The primary, longstanding goal of building codes for most buildings is to protect lives by reducing the likelihood of structural collapse in rare extreme events (i.e., Risk-targeted Maximum Considered Earthquakes), and to provide some level of property protection in more frequent events. Current codes generally do not consider design of buildings to explicitly provide functionality after a hazard event. Buildings that are designed per these codes may sustain extensive damage in a significant event requiring lengthy and costly repair or rebuilding, which in turn can lead to lack of providing the intended function or service. This loss of function negatively impacts sociocultural and economic functions of the community and can lead to temporary or permanent dislocation or required relocation of community members after an event [3]. Older existing buildings may have been built with no (or earlier and less effective) seismic design considerations, and may pose an even greater hazard than newer buildings.

In most cases, the current design of lifeline infrastructure systems (such as water, wastewater, power, gas and liquid fuels, transportation, and telecommunications) does not take into consideration the functionality of the system after an earthquake event. Unlike building design, the state of practice for design of different lifeline systems to provide critical services is more complex due to their interdependencies and need to address the broad spatially distributed networks of specialized components. It is expected that various lifelines systems may not be able to provide their intended services after an earthquake. Because buildings and their occupants

Sattar, Siamak; Johnson, Katherine. "Recommended Options for Improving the Functional Recovery of the Built Environment and Critical Infrastructure." Paper presented at 17 World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

depend on lifeline infrastructure systems, the lack of their critical service would undoubtedly impact the operation and ability of buildings to function and provide their intended services.

The U.S. public, particularly in large urban areas, will find the disruption of sociocultural and economic services not only distressing, but unacceptable for the timeframes that it would currently take to restore infrastructure or building functionality—which may be in the range of months to years depending upon the earthquake event. There is a need to improve the ability of buildings and infrastructure systems to continue functioning at full or acceptably reduced capacity post-event. Ensuring a more limited disruption of sociocultural and economic services will require significant effort to move beyond the current design paradigm. A new functional recovery performance objective would improve the performance of buildings and lifeline infrastructure systems, so that they are less likely to be negatively impacted and more likely to maintain a sufficient level of functionality or regain it in a timeframe acceptable to community members. By providing the basic intended function or service of various components of the built environment within an acceptable time following an earthquake, communities can mitigate and recover more quickly from earthquakes and reduce vulnerability and long-term negative consequences. Greater emphasis on functional recovery has the potential to reduce the cost and social, psychological, and health consequences for communities at risk of seismic events, and will in turn improve resilience across the nation.

3. Congressional Mandate

The federal government has recognized the above-mentioned problems and the need for improving the state of practice in design and retrofit of buildings and lifeline infrastructure systems. The 2018 reauthorization of the National Earthquake Hazards Reduction Program (NEHRP) (P.L. 115-307), included a new requirement for the National Institute of Standards and Technology (NIST) and the Federal Emergency Management Agency (FEMA) to convene a "committee of experts" and develop a "report of recommended options" to Congress for moving the built environment and critical infrastructure:

> *(a) ASSESSMENT AND RECOMMENDATIONS.—Not later than December 1, 2019, the Director of the National Institute of Standards and Technology and the Administrator of the Federal Emergency Management Agency shall jointly convene a committee of experts from Federal agencies, nongovernmental organizations, private sector entities, disaster management professional associations, engineering professional associations, and professional construction and homebuilding industry associations, to assess and recommend options for improving the built environment and critical infrastructure to reflect performance goals stated in terms of post-earthquake re-occupancy and functional recovery time.*
>
> *(b) REPORT TO CONGRESS.—Not later than June 30, 2020, the committee convened under paragraph (1) shall submit to the Committee on Commerce, Science, and Transportation, the Committee on Energy and Natural Resources, and the Committee on Homeland Security and Governmental Affairs of the Senate and the Committee on Science, Space, and Technology, the Committee on Natural Resources, and the Committee on Homeland Security of the House of Representatives a report on recommended options for improving the built environment and critical infrastructure to reflect performance goals stated in terms of post-earthquake re-occupancy and functional recovery time.*

In response to this mandate, NIST and FEMA convened a committee of experts with two components: the Project Technical Panel (PTP), which is responsible for developing the report, and the Project Review Panel (PRP), which is responsible for providing subject matter expertise peer review throughout the writing process. The report to Congress, hereafter known as NIST-FEMA report [4], addresses a breadth of mechanisms for driving change, including codes and standards, education, and planning and preparedness. In addition, five stakeholder workshops were held in St. Louis, MO; Salt Lake City, UT; Seattle, WA; San Francisco, CA; and Los Angeles, CA to gather broad national input to the NIST-FEMA report.

4. Functional Recovery

The NIST-FEMA report defines functional recovery as follows:

> *Functional recovery is a post-earthquake performance state in which a building or lifeline infrastructure system is maintained, or restored, to safely and adequately support the basic intended functions associated with the pre-earthquake use or occupancy of a building, or the pre-earthquake service level of a lifeline infrastructure system*.

Using concepts from performance-based earthquake engineering framework, design for functional recovery would involve the creation of a new functional recovery performance objective, defined as follows:

> *A functional recovery performance objective is functional recovery achieved within an acceptable time following a specified earthquake, where the acceptable time might differ for various building uses and occupancies, or lifeline infrastructure services*.

The notion of functional recovery supports community resilience goals by focusing on the design, construction, and retrofit of individual buildings and lifeline infrastructure systems. A functional recovery performance objective is one component of community resilience and can help a community to achieve resilience by enabling buildings and lifeline systems to recover their basic functions in a timely manner. The definition of *basic* intended functions may be somewhat less than full functionality, but more than what would be considered sufficient for simple re-occupancy of buildings, or temporary provision for critical lifelines services. The determination of basic, necessary, or critical services may require community context and information on the dependencies among various aspects of a community's built environment.

Although the ability of buildings and lifeline infrastructure systems to provide services depend on interactions among different systems, a functional recovery objective will be most efficient at the individual building or lifeline infrastructure system level. In current design philosophy and practice, buildings and lifeline infrastructure systems are designed separately using different codes and standards and are regulated by different jurisdictions and sectors. A new recovery-based design paradigm envisions separate but parallel functional recovery objectives applied to the design of individual buildings and lifeline infrastructure systems. These objectives must be coordinated between buildings and lifeline systems, for example, a building design will be informed by the expected performance of lifeline infrastructure systems, but not controlled by it. The separate functional recovery objectives for buildings and lifeline systems will prevent delay in progress due to complications in either sector and will be easier to implement as assets are constructed at different times by different stakeholders, using the then-applicable codes and standards. In this way coordinated and simultaneous development of functional recovery performance objectives is expected to expedite the progress toward the goal of reducing the time it takes for communities to recovery.

5. Design Based on Recovery Time (Beyond Safety)

As mentioned earlier, the primary objective of building codes for most buildings is achieving life safety. The International Building Code (IBC) currently categorizes building use or occupancy using a building's Risk Category as defined in the Table 1604.5 of the IBC [5]. The Risk Categories are developed based on the level of protection or risk to public safety. The Risk Categories do not represent the desired recovery time for different building uses, except at the highest risk category for essential facilities. To develop functional recovery performance objectives, the Risk Category concept can be extended to consider the desired recovery time for various building uses. This new categorization could be called the Recovery Category. In this new design paradigm, buildings and lifeline infrastructure systems would be designed to meet specific recovery time goals at a specified hazard level. The recovery categories can be determined based on the needed basic services from a building or lifeline infrastructure system, and the timeline these services are needed during response and recovery. The time required for recovery of function varies by the use, occupancy, and criticality

Sattar, Siamak; Johnson, Katherine. "Recommended Options for Improving the Functional Recovery of the Built Environment and Critical Infrastructure." Paper presented at 17 World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

of function that a building or lifeline infrastructure system provides. Not all services are needed immediately after an earthquake, nor are all services necessarily needed at the same time. Possible Recovery Categories for buildings and lifeline infrastructure systems are described in Table 1, which is used strictly as a means to illustrate the Recovery Category concept; the target recovery times and descriptions provided in Table 1 will be influenced and likely modified by future research.

Table 1: Recovery Categories [4]

Recovery Category (RC)	Target Recovery Time	Description
RC-4	Hours	Emergency Response - Basic services and systems needed for immediate response, rescue, and event stabilization to ensure emergency response activities can be undertaken
RC-3	Days	Short-Term - Basic services and systems needed at the initial stages of recovery
RC-2	Weeks	Intermediate-Term - Basic services and systems needed to restore neighborhoods and the workforce, and to care for historically underserved populations
RC-1	Months	Long-Term - Basic services and systems needed for restoring vitality to economy, sociocultural institutions, and physical infrastructure

In current performance-based design practice, buildings are designed to meet specific performance level(s) at specified hazard level(s). In current practice, most buildings across the U.S. required to meet seismic provisions are designed to provide life-safety protection at a "design-level" hazard event. The design-level earthquake is an earthquake with a statistical likelihood of occurring once in every 300 to 700 years (also called a return period), at a particular location throughout the country. The seismic hazard level tentatively considered for functional recovery considerations in this paper, i.e., Table 1, is also taken to be the design-level event. Alternatively, scenario-based events may be more appropriate for the basis of design for locations with well-defined fault mechanisms. Future research is needed to inform selection of an appropriate hazard level for the design of functional recovery.

6. Recommended Options for Improving Built Environment and Critical Infrastructure

The NIST-FEMA report presents nine recommendations across four areas of emphasis including developing a national recovery framework, improving the built environment, improving planning, and raising awareness and understanding of potential earthquake impacts on the built environment and lifeline infrastructure systems. This section summarizes these nine recommendations. Further information regarding different options for implementation of each recommendation may be found in the NIST-FEMA report. Please note that the nine recommendations discussed here are preliminary recommendations and are subject to modification in the final draft of the NIST-FEMA report.

6.1 Develop a National Functional Recovery Framework

One of the fundamental steps identified in the NIST-FEMA report to support the development and implementation of functional recovery performance objectives is developing a functional recovery framework for buildings and lifeline infrastructure systems. This framework would address the multidisciplinary aspects of functional recovery. The framework needs to incorporate policy, social science, and engineering aspects of the issue. The framework would identify acceptable recovery times for different building uses and lifeline infrastructure services and would also address the required provisions to achieve the desired performance goals. A minimum standard is recommended for consistency across the nation, while still allowing local jurisdictions to exceed the minimum recommendations according to their priorities and distinctive challenges. The national framework should determine the functions that are critical to recovery as well as their desired timeline. The framework should also consider mechanisms to support coordination between the desired recovery times of buildings and lifeline infrastructure systems, as well as the costs and benefits associated with selecting particular hazard levels or recovery times. Future research will be needed to provide the information required for developing the national framework.

6.2 Design New Buildings to Meet Recovery-based Performance

Buildings designed according to current codes and standards may experience significant damage during a design-level earthquake that can hinder the intended function of the building. A cornerstone to all of the options for achieving functional recovery goals is to design new buildings to meet recovery-based design objectives. In this new design paradigm, in addition to designing for life-safety objectives, new buildings will also be designed to satisfy a specific recovery time after a design-level event. One of the first steps in this process is to benchmark the recovery time that current buildings codes and standards deliver. It is possible that some current building uses may already meet the desired recovery time. If the benchmarking results identify a need for reducing the desired recovery time, two alternative approaches may be pursued. In the first approach, the design requirement for a higher Risk Category building can be applied to a broader class of new structures. In an alternative approach, new buildings could be designed using new codes and standards with design criteria developed to achieve re-occupancy and recovery of function in an acceptable timeframe. Regardless of the chosen approach, the implementation of this recommendation may be achieved through either mandatory or voluntary mechanisms, and at the national, state, or local levels. Both mandatory and voluntary mechanisms are associated with implementation pros and cons that will require careful evaluation.

6.3 Retrofit Existing Buildings to Meet Recovery-Based Objectives

Enhancing the performance of existing buildings is a critical aspect of improving community resilience since existing buildings comprise the majority of the building stock and pose the greatest threat to the community. Existing buildings are more challenging to address than new buildings, as improving the performance of this building group is constrained by various factors including the technical feasibility of achieving higher performance goals, as well as the costs associated with retrofit. There is also a concern that aiming towards higher functional recovery targets (i.e., relatively short target recovery times), for retrofitting existing buildings may adversely impact safety-targeted retrofit actions. This would be an unintended negative consequence of aiming for greater resilience. One way to manage challenges related to existing buildings is to adopt lower re-occupancy or functional recovery goals than for new buildings. Such an approach will mitigate the greatest risks associated with marginally deficient buildings. In addition, retrofit programs can be paired with pre-event planning such as developing re-occupancy plans and relocation of critical uses to enhance the effectiveness of the retrofit programs. Regardless of the target recovery-based performance goal for existing building retrofits, local jurisdictions need to identify buildings that require re-occupancy and functional recovery design, as well as actionable triggers for retrofit and appropriate requirements. Similar to the design of new buildings, implementation of existing building retrofits can be done through mandatory or voluntary approaches, using national, state, or local design criteria. Both mandatory and voluntary mechanisms are associated with implementation pros and cons that will require careful evaluation.

6.4 Design, Upgrade, and Maintain Lifeline Infrastructure Systems to Meet Recovery-based Performance Objectives

Lifeline infrastructure systems have numerous operational requirements as well as regulatory environments. Most regulations for lifeline systems focus primarily on public health, as well as safe and reliable operations. These regulations are not currently intended to enable the provision of services in a specific timeframe after most hazard events, including earthquakes. Although there are multiple manuals and guidelines for design of the components of lifeline systems, design criteria are inconsistent among systems, and most of them do not incorporate seismic design. Lifelines infrastructure systems are vital components of the built environment and community recovery highly depends on the recovery of the services they provide. Therefore, there is a critical need for a shift in the design paradigm of lifeline infrastructure systems from protecting lives and property to focus on recovery of function after a hazard event within an acceptable timeframe.

To ensure consistent design and operations throughout the systems and among various owners and operators, national-level seismic design guidelines, standards, and codes are needed. These guidelines, standards, and codes would be based on functional recovery performance objectives. To develop the functional recovery design paradigm for lifeline systems, clear guidance and multiple implementation and support tools are needed. The development of codes, standards, guidelines, and tools will require additional research. Implementing the Earthquake Resistant Lifelines: NEHRP Research, Development and Implementation Roadmap [6] can serve as a stepping stone for this work.

Due to the interconnected, complex, and interdependent nature of lifeline infrastructure systems, coordinated efforts across various stakeholders are needed to develop and implement coherent and consistent performance goals for lifeline infrastructure systems. Continued support for the development of state or regional lifeline councils could significantly help engagement from different lifeline infrastructure owners and operators and is essential to implementing the framework at local levels. Besides the above-mentioned design and implementation activities, seismic resilience plans for each lifeline infrastructure system need to be developed. These plans should focus on improving (1) pre-earthquake integrated asset management plan to address aging/vulnerable components to enhance system-level resilience and (2) post-earthquake disaster recovery plans for rapidly repairing and recovering the systems. The re-establishment of the national program, as part of NEHRP, to advance the engineering of lifeline infrastructure systems could significantly assist with the leadership, management, and coordination of cross-country efforts.

6.5 Develop and Implement Plans Needed to Facilitate Functional Recovery

While codes and standards are necessary to achieve functional recovery goals, they are not sufficient in themselves. In addition, robust planning activities that enable the success of a functional recovery objective are needed. Planning is an essential step towards meeting functional recovery goals. Effective plans engage relevant and representative stakeholders in dialogue around mutually-agreed upon goals and objectives, strategies and tactics. Engaging diverse stakeholders, by educating them, getting buy-in, and feedback from them about functional recovery actions, will increase the chances of successful implementation of the functional recovery framework. Planning for functional recovery can be incorporated into all types of ongoing and future community plans.

One of the first steps related to planning efforts is to adjust the language and tools in the existing mitigation plans such as FEMA's local mitigation plan [7], that spells out requirements for communities to be eligible for FEMA post-earthquake aid for public and non-profit facilities. Similarly, functional recovery can be integrated into the community resilience activities undertaken by chief resilience officers, emergency managers, community development professionals, and other similar representatives. Moreover, state and local government can play an important role in achieving functional recovery goals by considering improvements to the development and implementation of regulatory incentives for mitigation plans, supporting planning for swift re-housing efforts after an earthquake, and the development of protective measures that can help communities to recover quickly and maintain their populations.

6.6 Perform Rapid Building Inspections and Evaluations to Facilitate Functional Recovery

Assessment of buildings' performance after an earthquake is essential to evaluate whether the they are safe to occupy and also to determine their post-hazard functionality level. Timely inspection and evaluation of buildings after an earthquake plays an important role in expediting the recovery process. There are multiple opportunities for improving the state of practice concerning inspection and evaluation of buildings ranging from development of technical guidelines to programs and policies that can support re-occupancy and functional recovery efforts [8], such as a pre-arranged plan to provision building inspectors.

From a technical standpoint, inspection guidelines need further development to effectively incorporate recovery-based assessment tools as current inspection guidelines primarily focus on the safety of buildings. Recent advances in remote sensing technologies make it possible to use seismic instrumentation to expedite the assessment and recovery of buildings and lifeline infrastructure systems. Improvements and development of these technologies are needed and may be essential for certain types of buildings. In addition, development of alternative standards for temporary habitability of buildings in post-earthquake scenarios may mitigate current stringent requirements for evacuation of buildings during the repair process. Another factor that can prevent unnecessary evacuation of residents is developing protocols for establishing safety cordons around damaged buildings that consider their associated risk with damaged buildings in conjunction with the disruption in the recovery process.

From the policy perspective, there is a need to enact policies and procedures that facilitate post-earthquake safety inspection of buildings by state and local government. For example, development and implementation of programs for funding and sharing local jurisdiction and county staff for earthquake inspection and recovery programs after an earthquake can significantly expedite the inspections and tagging process.

6.7 Explore Financial Resources to Facilitate Functional Recovery

Speedy access to financial capital after an earthquake event plays a key role in expediting recovery. There are currently different post-disaster funding mechanisms including federal programs, insurance, and loans; one common factor among these mechanisms is the slow process of administration that can significantly delay the recovery process. Improving the speed of access to resources plays an important role in expediting the recovery of buildings and lifeline infrastructure systems meeting functional recovery objectives; this can occur by modifying existing or developing new financial programs. Further coordination of existing federal programs to enable quicker access to and distribution of funds to local jurisdictions in post-event situations is also needed. One example related to access to federal financial support would be developing a Federal Case Management System and single application form, where one application can be submitted by the building owner for all applicable federal programs for disaster assistance. Consideration can be given to improving the availability of affordable housing loans after earthquakes as well as improving grant programs to expedite access to financial resources. Federal agencies may also encourage the use of parametric insurance for homes and businesses.

In addition, consideration can be given to improving the access to quick funds after an earthquake event through natural hazards insurance programs, and particularly by additional work to supporting the development of a fiscally viable and affordable earthquake insurance program. Other programs to facilitate access to post-disaster funds by private individuals, such as disaster accounts, or pre-arranged repair loans, could significantly help building owners more quickly begin the journey through repair and restoration of their buildings.

6.8 Educate Building Owners, Tenants, and Customers/Users about Building Performance Expectations and Enable Action that Will Lead to Functional Recovery

Public acceptance and input are important components of any functional recovery design and implementation. Effective public outreach requires developing and implementing appropriate educational materials, incorporating risk communication methods, and establishing ongoing engagement mechanisms. The public should be educated on the current safety-based target of the building code and its consequences on the community and its built environment in an event of an earthquake. This educational effort may help support the need for an enhanced performance target. In addition, educating building owners, tenants, and customers on the benefits of the functional recovery design may positively influence their willingness to enact functional recovery concepts. Additional work is needed to help inform the users and owners of buildings and lifeline infrastructure systems about the significant shifts required to achieve functional recovery objectives and how they are fundamentally distinct from current practice. Key components of the educational effort should include enhancing public understanding of their level of risk from seismic activity and what their buildings or lifeline infrastructure systems may or may not be able to provide given their current state. Stakeholders should also be educated on mitigation and preparedness strategies that can improve the recovery of function after an earthquake. Federal agencies could play a significant role in this effort not only by educating the public, but also by creating and promoting a nationwide seismic continuity program for all building uses to owners and tenants to address their respective unique situations. Consideration should be given to improving the continuity programs to identify effective mitigation and preparedness strategies for resuming functions post-earthquake in a timely manner.

6.9 Enhance Outreach and Continuing Education Efforts for Building Industry Professionals and their Associations

The functional recovery efforts will be more successful if adopted by many individuals, organizations, and jurisdictions. Recruiting and maintaining a workforce knowledgeable about functional recovery and implementation methods will be crucial to ensure a common understanding across the building professions and government agencies. The workforce includes engineers, architects, contractors, code officials, etc. The professional associations of the building industry are key players in moving the nation toward functional recovery because they can reach out to significant numbers of building design and construction professionals. Multiple activities can be undertaken to enhance outreach and continuing education for building industry professionals. For example, functional recovery concepts along with a discussion of the societal benefits of functional recovery could be added to the codes of ethics for different building industry associations. Similarly, the continuing education requirements for professional licensing could incorporate the functional recovery concept.

Designing buildings to functional recovery performance objectives will be a notable shift from current standards of practice for the engineering and architectural fields. Training programs will be essential to introduce skilled professionals to new concepts and to educate them on new codes, regulations, and inspection protocols.

7. Workshops

Functional recovery inherently incorporates risk tolerance, community preferences, and societal values. As a result, it is critical to gather feedback from stakeholders on developing the functional recovery framework and prioritizing options for improving functional recovery time. As part of the congressionally mandated effort, five stakeholder workshops were held across the U.S. to gather input from a broad range of community leaders and subject matter experts throughout the community on concepts that will inform the functional recovery framework and the NIST-FEMA report. The workshops were conducted in St. Louis, Salt Lake City, Seattle, San Francisco, and Los Angeles. The workshop participants represented a broad range of stakeholders and subject matter experts, including local officials, private consultants, structural engineers, social scientists, utility and lifeline system representatives, and others. The workshop collected information on three main topics: 1) the time-frame for recovery of difference components of the built environment that support various social functions, 2) attributes for evaluating and assessing options for improving functional recovery time, and 3) trade-offs among different attributes to inform evaluation of options for improving functional recovery time.

[The workshops recently concluded. Assessing results of the workshops is currently in progress. The final manuscript and presentation will include key findings from the workshops].

8. Summary

Buildings and lifeline infrastructure systems designed per current codes and standards may sustain extensive damage during an earthquake event. The widespread damage is likely to hinder the service that buildings and lifeline infrastructure systems provide to support the sociocultural and economic functions of a community [3]. The required time for regaining function of various components of the built environment is either likely to be extensive, or not well understood and difficult to estimate. A post-earthquake state in which people may not have access to their jobs, schools, housing, and other services can produce sociocultural and economic consequences that can lead to temporary or permanent displacement of a community's population. The U.S. federal government has recognized the need to extend the target of risk mitigation for the built environment from life safety to include timely recovery of function. The functional recovery design concept is proposed as a means to achieve the target performance of timely recovery of function after an earthquake. This paper summarizes the effort undertaken by NIST and FEMA in response to a Congressional mandate to recommend *"options for improving the built environment and critical infrastructure to reflect performance goals stated in terms of post-earthquake re-occupancy and functional recovery time"*. The paper provides an overview on nine recommendations across four areas of emphasis related to developing a national recovery framework, improving the built environment, improving planning and expediting response and recovery, and raising awareness and understanding of potential earthquake impacts on the built environment. Future research and consensus-based decision making are needed to prioritize these options; attributes such as costs, benefits, impact, feasibility, and timeline are likely candidates for evaluation criteria. Achieving functional recovery across a community requires a multi-faceted approach that includes parallel efforts on aspects of: design of new buildings; retrofit of existing buildings; lifeline infrastructure systems; planning, outreach, and education. Developing and implementing functional recovery objectives represents a significant shift in the design philosophy for buildings and lifeline systems that demands a multi-disciplinary perspective and engagement from a broad range of community stakeholders, but this can happen within mechanisms currently in place for codes and standards development, policy development, and community resilience planning. Achieving greater seismic resilience through functional recovery goals will benefit not only local communities; the entire nation will benefit. This work can be accomplished through the strong partnerships and interactions across traditionally disparate sectors that have already initiated efforts on functional recovery.

9. Acknowledgements

The authors express their gratitude to members of the Project Technical Panel members: David Bonowitz, Mary Comerio, Greg Deierlein, Susan Dowty, Gary Ehrlich, Ron Hamburger, Ron Larsen, Rachel Minnery, Sissy Nikolaou, Chris Poland, Jonathan Siu, Jeffrey Soulages, and Kent Yu, as well as Project Review Panel members: Marissa Aho, John M. Bozeman, Dan Eschenasy, Robert Ezelle, William Holmes, James Kendra, Thomas O'Rourke, Marios Panagiotou, Keith Porter, and Lakisha Woods. The authors also thank the Applied Technology Council (ATC) for supporting the Project Technical Panel, the Science Technology Policy Institute (STPI) for organizing and facilitating the stakeholder workshops, and all of the participants in the stakeholder workshops.

10. References

[1] Jaiswal, K., Peterson, M., Rukstales, K., and Leith, W. (2015): Earthquake Shaking Hazard Estimates and Exposure Changes in the Conterminous United States, *Earthquake Spectra*, vol. 31, pp. S201–S220.

[2] FEMA (2017): Hazus Estimated Annualized Earthquake Losses for the United States, FEMA P-366, Federal Emergency Management Agency, Washington, D.C.

[3] Arendth, L. and Alesch, J. (2014): *Long-term Community Recovery from Natural Disasters*. Boca Raton, Florida: CRC Press.

[4] NIST-FEMA (2020): Recommended Options for Improving the Built Environment for Post-Earthquake Reoccupancy and Functional Recovery Time, National Institute of Standards and Technology and Federal Emergency Management Agency, FEMA P-XXXX/NIST SP-XXXX, (in-progress).

[5] International Code Council (2017): *2018 International Building Code*. Washington, D.C.

[6] NIST (2014): Earthquake Resilient Lifelines: NEHRP Research, Development, and Implementation Roadmap, NIST GCR 14-917-33, National Institute of Standards and Technology, Gaithersburg, Maryland.

[7] FEMA (2011): Local Mitigation Plan Review Guide, Federal Emergency Management Agency, Washington, D.C.

[8] FEMA (2019): Post-disaster Building Safety Evaluation Guidance, Report on the Current State of Practice, including Recommendations Related to Structural and Nonstructural Safety and Habitability, FEMA P-2055, Federal Emergency Management Agency, Washington, D.C.

Sattar, Siamak; Johnson, Katherine. "Recommended Options for Improving the Functional Recovery of the Built Environment and Critical Infrastructure." Paper presented at 17 World Conference on Earthquake Engineering, Sendai, JP. September 13, 2020 - September 18, 2020.

Zero-Compensation Josephson Arbitrary Waveform Synthesizer at 1.33 V

Nathan E. Flowers-Jacobs[1], Akim A. Babenko[1], Anna E. Fox[1],
Justus A. Brevik[1], Paul D. Dresselhaus[1], and Samuel P. Benz[1]
[1]National Institute of Standards and Technology, Boulder, CO 80305, USA
nathan.flowers-jacobs@nist.gov

***Abstract*—This paper describes the generation of a quantum-based rms output voltage of 1.332 V using an ac-coupled Josephson Arbitrary Waveform Synthesizer (JAWS) without any low-frequency compensation current biases, that is, in a 'zero-compensation' (ZC) mode. Low-frequency bias currents result in unwanted error voltages that increase with frequency. The ZC technique minimizes these currents and removes the resultant error voltages.**

***Index Terms*—Digital-analog conversion, Josephson junction arrays, signal synthesis, superconducting integrated circuits, voltage measurement.**

I. Introduction and System Description

Josephson Arbitrary Waveform Synthesizers (JAWS) use current-pulse-biased arrays of Josephson junctions (JJs) to generate quantum-based voltages. When operating correctly, each current pulse causes each JJ to generate a voltage pulse with an integrated area $h/(2e)$ that is dependent only on fundamental constants: the electron charge e and the Planck constant h. By controlling the pulse-timing, JAWS systems are used to create spectrally pure low-frequency sine waves for ac voltage metrology along with other arbitrary waveforms [1].

To create large output voltages for metrology, many JJs and narrow pulses with a fast repetition rate are required. In this paper, we use a cryocooled system at 4.5 K with 204 960 JJs across two separate chips with four independent high-speed current pulse biases [2]. The JJs have critical currents between 5.9 mA and 6.5 mA and are placed in series arrays along the center-conductors of 16 coplanar waveguides.

There are a number of sources of error that can impact the JAWS output voltage, including the transfer function of the output leads, leakage resistance and capacitance, feedthrough of the bias signal, and inductive voltage errors [3]. In this paper, we have increased the maximum output voltage that can be generated using a 'zero-compensation' (ZC) technique [4] that reduces the last two sources of error. Both of these errors are caused by bias signals at low-frequencies that create undesired error voltages at the DUT. The ZC technique uses a current bias that, ideally, has negligible current at low-frequencies and thus removes the source of the error voltages. However, this technique depends on the JJ nonlinearity to convert an input current pulse stream with no low-frequency content into a quantum-based output voltage pulse stream with calculable low-frequency content.

By optimizing the algorithm used to generate the bias waveform, the microwave bias electronics, and the shape of the current bias pulses, we create a ZC waveform with an rms magnitude of 1.332 V at 10 kHz (Fig. 1).

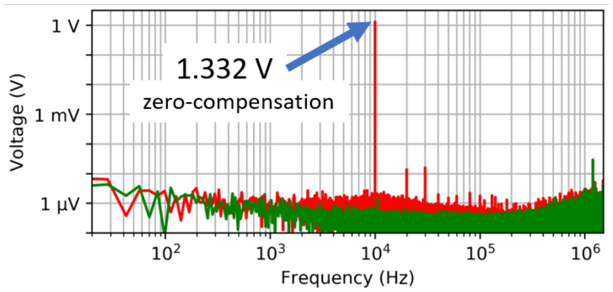


Fig. 1. Spectrum of a quantum-based 'zero-compensation' JAWS rms output voltage of 1.332 V at 10 kHz (red, harmonics are due to the digitizer nonlinearity) and background with no bias applied to the JJ arrays (green). The spectrum is sampled at 3.75 megasamples per second.

II. Current Bias Waveform

We use an 8-bit 57.6 gigasample-per-second arbitrary waveform generator (AWG) to create the current pulse bias. The waveform stored in the memory of the AWG is generated using a two-step procedure. First, a delta-sigma algorithm is used to determine the desired timing of the pulses from the JJs. The algorithm uses a three-level output corresponding to either a positive JJ pulse, a negative JJ pulse, or no JJ pulse at each clock step with a clock rate of 14.4 GHz. We also constrain the algorithm so that pulses with opposite signs are separated by at least 50 clock cycles and pulses with the same sign are separated by at least 3 clock cycles.

Second, we digitally high-pass filter the AWG waveform to remove all of the low frequency content below 100 MHz. This algorithm uses the multi-bit nature of the AWG and typically results in a reduction in the low-frequency current content below 100 MHz by at least four orders of magnitude when generating large output voltages.

The AWG is followed by a high-power, wide-bandwidth amplifier. The amplifier non-linearity when operating near saturation mixes down part of the large power components in the pulse-bias around 14.4 GHz, which creates a low-frequency signal at the synthesis frequency after the amplifier. We employ a two-pole high-pass filter between the amplifier and the cryostat to reduce this low-frequency component of the bias.

Finally, we also use the AWG finite-impulse-response (FIR) filter to optimize the shape of the current bias pulses. We start by surrounding each pulse with opposite sign half-pulses to create a composite pulse-like object which further reduces

Flowers-Jacobs, Nathan; Babenko, Akim; Fox, Anna; Brevik, Justus; Dresselhaus, Paul; Benz, Samuel P. "Zero-Compensation Josephson Arbitrary Waveform Synthesizer at 1.33 V." Paper presented at 2020 Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

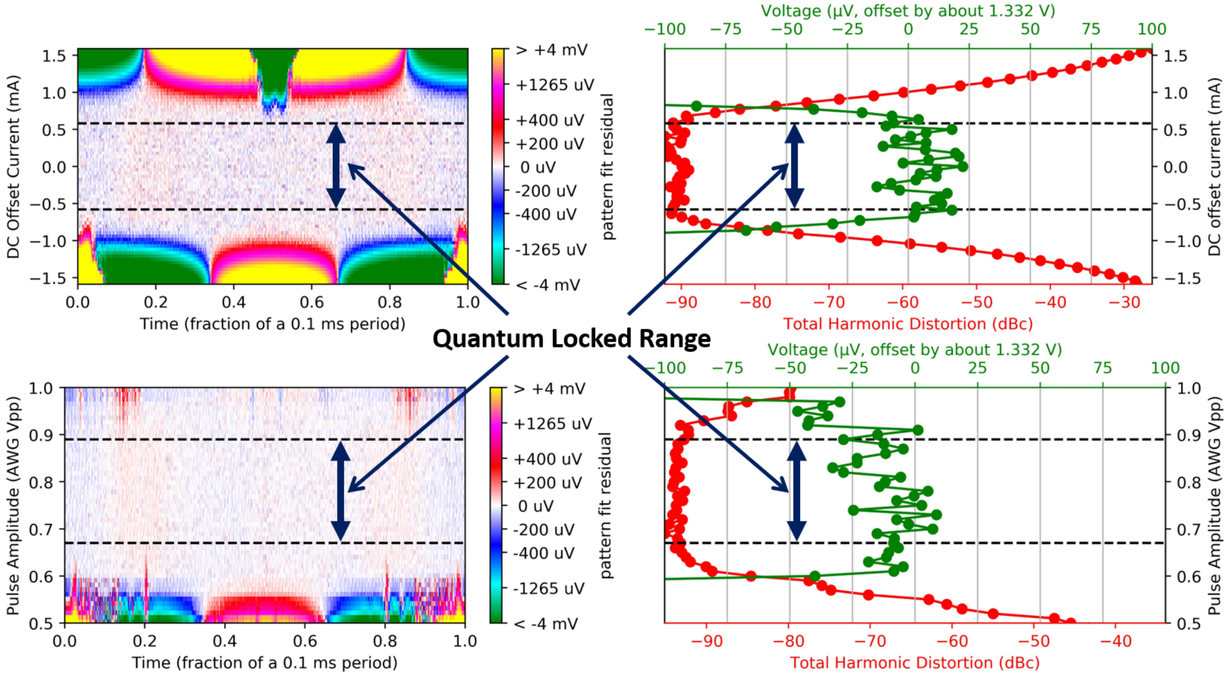


Fig. 2. Density plots showing the residual of a fit to a sinusoid as a function of time and bias dither (left) and the generated voltage (green) and THD (red) versus bias dither (right). The dithered bias is either a dc current offset on all JJ arrays (top) or the pulse amplitude applied to a quarter of the JJ arrays (bottom). The quantum locked range is between the dashed black lines.

the low-frequency content produced by the AWG, though the detailed shape of the pulse is then modified to optimize the quantum-locked range (see next section). The width of this pulse-like object sets a limit on the pulse repetition rate, which in turn places a limit on the maximum output voltage.

III. Quantum Locked Range

Using the techniques discussed in the previous section, we were able to generate a quantum-based output rms voltage without compensation of 1.332 V at 10 kHz (Fig. 1). However, a clean spectrum is not proof of quantum-based behaviour. We also need to confirm that every JJ generates a single output voltage pulse for every input current bias pulse over a range of bias parameters. We call this range the ‘quantum locked range’ (QLR) for the dithered bias parameter. For ZC waveforms, the two most relevant bias parameters are the dc current and the amplitude of the current pulses.

In Fig. 2 (left) we plot the residuals of a fit to the 10 kHz sinusoid as a function of waveform period (x-axis) and dc offset current through all of the JJs (y-axis, top) or the pulse amplitude applied to a quarter of the JJs (y-axis, bottom); the response to changing other pulse amplitudes is similar. From the same data, we also extract and plot (right) the magnitude of the fundamental and the total harmonic distortion (THD). We observe a clear dc current offset QLR with a width greater than 1.1 mA over which the magnitude and THD are constant. The similar pulse amplitude QLR is 0.2 V or 20 % of the tuning range of the AWG. In this case, the units of the QLR do not directly correspond to current through the JJ arrays.

At CPEM 2020 we intend to expand this preliminary data beyond measurements at a single frequency and present a more comprehensive comparison between the output voltages generated by the different current pulse bias channels.

IV. Conclusion

In conclusion, we were able to generate a quantum-based rms output voltage of 1.332 V without compensation using a cryo-cooled JAWS system. This ‘zero-compensation’ bias mode results in smaller voltage errors in comparison to the more typical JAWS bias mode that requires low-frequency bias currents.

Acknowledgment

We thank Robert Schwall for cryocooler development, and the Boulder Microfabrication Facility for fabrication support.

References

[1] A. Rüfenacht, N. E. Flowers-Jacobs, and S. P. Benz, “Impact of the latest generation of Josephson voltage standards in ac and dc electric metrology,” *Metrologia*, vol. 55, no. 5, pp. S152–S173, Oct 2018.

[2] N. E. Flowers-Jacobs, A. E. Fox, P. D. Dresselhaus, R. E. Schwall, and S. P. Benz, “Two-Volt Josephson Arbitrary Waveform Synthesizer Using Wilkinson Dividers,” *IEEE Transactions on Applied Superconductivity*, vol. 26, no. 6, pp. 1–7, Sep 2016.

[3] J. M. Underwood, “Uncertainty analysis for ac–dc difference measurements with the AC Josephson voltage standard,” *Metrologia*, vol. 56, no. 1, p. 015012, Feb 2019.

[4] J. A. Brevik, N. E. Flowers-Jacobs, A. E. Fox, E. B. Golden, P. D. Dresselhaus, and S. P. Benz, “Josephson Arbitrary Waveform Synthesis With Multilevel Pulse Biasing,” *IEEE Transactions on Applied Superconductivity*, vol. 27, no. 3, pp. 1–7, Apr 2017.

STRUCTURE IGNITION MECHANISMS IN WILDLAND-URBAN INTERFACE (WUI) FIRES AND ASSOCIATED CALIFORNIA STATE FIRE MARSHAL STANDARD TEST METHODS

Samuel L. Manzello
National Institute of Standards and Technology (NIST), USA

1. INTRODUCTION

Wildland fires that spread into communities, known as wildland-urban interface (WUI) fires, are a major global problem. In the USA, a series of very destructive WUI fires have occurred over the past several decades, mainly in California. The most significant of these were the 2018 Northern California fires that destroyed more than 18,800 structures and resulted in scores of fatalities. WUI fires are not limited to the USA, as South Korea experienced significant fires in 2019. Australia, like the USA, has a long history of destructive WUI fires, and the 2020 Australian WUI fires were worldwide news for months.

More than a decade ago, the state of California in the USA identified the need to develop an approach to harden communities to WUI fires exposure. The term harden simply indicates to make infrastructure in communities more ignition resistant. The premise has its roots in the development of standards and codes developed to mitigate urban fire disasters that were observed in the USA, such as the 1871 Great Chicago Fire or 1904 Baltimore Fire. The urban fire codes and standards provide the basis for fire resistant construction in many countries throughout the world [1]. In the USA, the concept of WUI fire building codes and standards is far newer, due to the more recent WUI fire problem in this country. Developing test standards for outdoor fire exposures presents significant challenges.

The objective of this study is to provide an overview of WUI structure ignition mechanisms and the associated standard test methods developed by the State Fire Marshal (SFM) of California [2]. The paper closes with a short discussion on the need for improved WUI standard test methods.

2. WUI FIRE STRUCTURE IGNITION MECHANISMS

The main ignition mechanisms of structures from WUI fire exposures are due to direct flaming combustion contact, radiant heat, and firebrand exposure. Direct flaming combustion contact refers to the situation where a structural component is in direct contact with flaming combustion from an adjacent combusting fuel source. In WUI fires, this could be ornamental vegetation, such as mulch, shrubs, or trees, or other fuel types, such as a burning vehicle or a neighboring structure [3].

Radiant heat is a form of electromagnetic radiation that is emitted from sufficiently hot materials. Due to the combustion of vegetative and structural fuels in WUI fires, any fuel type in close proximity to these combustion processes will experience radiant heat that may lead to ignition of these adjacent fuels.

Firebrands are generated from the combustion of both vegetative and structural fuels in WUI fires. The most important hazard is if the deposited firebrands have sufficient energy to ignite these adjacent fuel sources. In most cases, the firebrands themselves are in a state of glowing combustion since it is difficult to sustain flaming combustion as they are transported through the atmospheric boundary layer. Once firebrands land on fuel beds, the initial ignition mechanism is that of smoldering combustion. If a glowing firebrand is able to provide enough energy, self-sustained smoldering combustion may occur in the fuel bed, and under the influence of an applied wind field, the smoldering combustion reaction may transition to a flaming combustion reaction. Firebrand exposure is a dominant mechanism to structure ignition in WUI fires [4].

3. CALIFORNIA STATE FIRE MARSHAL (SFM) TEST STANDARDS

Based on post-fire studies of ignition vulnerabilities in WUI fires, standard test methods were devised by the office of the State Fire Marshal (SFM) in California to address ignition vulnerabilities to exterior walls, exterior windows, horizontal projections such as eaves, decking assemblies, and the use of ignition resistant materials. It is important to note that the exposure conditions used in these test methods are best guess estimates of what exposure conditions would be in a WUI fire and were developed with the best available information at that time.

Manzello, Samuel. "Structure Ignition Mechanisms in Wildland-Urban Interface (WUI) Fire and Associated California State Fire Marshal Standard Test Methods." Paper presented at Japan Association for Fire Science and Engineering, Hirosaki, JP. May 28, 2020 - May 29, 2020.

3.1SFM-12-7A-1 EXTERIOR WALL SIDING AND SHEATHING

This standard test method makes use of a burner to expose a wall to a 10 min 150 kW fire exposure. The test attempts to simulate a scenario where an exposure of flame impingement occurs as a result of ignition of plants, trash, a deck, or other combustible materials beside the wall. The test method does not apply to a scenario where the building is exposed to a large radiation source for a long duration of time, such as burning of an adjacent structure.

3.2 SFM-12-7A-2 EXTERIOR WINDOWS

To assess the performance of windows exposed to direct flames, this test method uses a 150 kW burner under a target window. The specimen is exposed to the flame for 8 min. This test method is not representative of a radiation exposure from a WUI fire.

3.3 SFM-12-7A-3 HORIZONTAL PROJECTION UNDERSIDE (EAVES)

This standard test exposes a projection to a flame of 300 kW for 10 min. The sample is then observed for another 30 min to monitor the existence of glowing or flaming combustion on the unexposed side of the specimen.

3.4 SFM-12-7A-4 DECKING

In part A, the test method exposes decking assembly samples to direct flame exposure of 80 kW for 3 min. The sample fails if there is runaway combustion, structural collapse, or flaming dripping materials. The test procedure requires observation of the sample for 40 min after the flame exposure.

In part B, a standard burning firebrand (Class A crib from ASTM E108 [5] to represent a firebrand) is placed on the deck while a fan blows an air flow of approximately 5.4 m/s over the specimen and the sample is observed for 40 min for signs of sustained flaming or falling firebrands. In all cases, the samples need to be exposed to conditions of accelerated aging or weathering to create a more realistic representation of the actual decking assembly in the test.

3.5 SFM-12-7A-5 IGNITION RESISTANT MATERIAL

Any material designated as ignition resistant must pass a 30 min ASTM E84 [6] test. ASTM E84 was not developed by the SFM of California but is a legacy standard test method, also known as the Steiner Tunnel test method.

4. NEED FOR IMPROVED TEST METHODS

While these test methods described above attempt to harden communities to WUI fire exposures, these standards remain largely unproven to actually mitigate WUI fire spread and structure ignition processes. Post-fire investigations conducted by NIST and elsewhere have indicated that most structures in WUI fires are destroyed from firebrand exposure [4].

The development of the NIST Firebrand Generator (NIST Dragon) coupled with the Fire Research Wind Tunnel Facility (FRWTF) at the Building Research Institute (BRI) in Japan have led to improved regulations in California. As an example, a recently developed test method, ASTM E2886 [7], has been implemented to resist firebrand entry into building vents. This standard is listed as part of California’s requirements for firebrand resistant building vents in WUI hazard areas and BRI/NIST experiments form the scientific basis for this test method. While there is no special SFM test method for roofing assemblies, classification is based on the legacy ASTM E108 [5] roof test method. Based on BRI/NIST experiments that demonstrated firebrands may penetrate under roof tiles, California requires the addition of a non-combustible cap-sheet placed under roof tiles to mitigate firebrand ignition. Fencing assemblies are known to be vulnerable to firebrand exposure, but these have yet to be addressed in California [2]. Recent firebrand shower experiments have also demonstrated the deficiencies of SFM-12-7A4 part B, the decking test method.

5. SUMMARY

WUI fire structure ignition mechanisms and associated standard test methods in use by the state of California were discussed. To be able to lessen the destruction from WUI fire exposures requires the development and implementation of improved scientifically based standards and codes.

6. REFERENCES

[1] N.P. Bryner, Building Codes and Standards for New Construction, in: S.L. Manzello (Ed.), Encyclopedia of Wildfires and Wildland-Urban Interface (WUI) Fires.

[2] 2019 California Building Code, Title 24, Part1, Chapter 7A Materials and Construction Methods for Exterior Wildfire Exposure.

[3] E. Pastor, Direct Flame Contact, in: S.L. Manzello (Ed.), Encyclopedia of Wildfires and Wildland-Urban Interface (WUI) Fires.

[4] R. Blanchi, A. Maranghides, and J. England, Lessons Learnt from Post-Fire Investigations and Surveys, in: S.L. Manzello (Ed.), Encyclopedia of Wildfires and Wildland-Urban Interface (WUI) Fires.

[5] ASTM E108-00, Standard Test Methods for Fire Tests of Roof Coverings.

[6] ASTM E84, Standard Test Method for Burning Characteristics of Building Materials.

[7] ASTM E2886/E2886M-14, Evaluating the Ability of Exterior Vents to Resist the Entry of Embers and Direct Flame Impingement.

Toward Ultra-fast and Ultra-low Power Switching in Perpendicular Magnetic Tunnel Junctions

Daniel B. Gopman
Materials Science & Engineering Division
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
daniel.gopman@nist.gov

Weigang Wang
Department of Physics
University of Arizona
Tucson, AZ 85721, USA

William Taylor
GLOBALFOUNDRIES
Malta, NY 12020, USA

Jian-Ping Wang
Tony Low
Department of Electrical & Computer Engineering
University of Minnesota
Minneapolis, MN, USA

Abstract— **Magnetic random access memory (MRAM) based on perpendicular magnetic tunnel junctions (pMTJs) is one of the core building blocks for beyond-CMOS technologies. Their inherent non-volatility, rad-hardness and endurance (10^{16}) makes pMTJs extremely competitive for computation-in-memory and other DoD applications where security and ruggedness are of the utmost vitality. To deliver on the potential of MRAM for computation-in-memory, reductions in the energy- and delay characteristics of pMTJs must be demonstrated. Based on interfacial- and bulk perpendicular magnetic anisotropy materials, we demonstrate two novel perpendicular synthetic antiferromagnet (p-SAF) designs for ultra-fast and ultra-low power switching performance: one using interfacial PMA materials and one using bulk PMA materials. Our stacks are compatible with or close to the existing p-MTJ stack and fabrication process, which will make the technology transition to back-end-of-line semiconductor process practical within a 5-10 year time frame.**

Keywords—MRAM; beyond-CMOS; spintronics; rad-hard

I. Introduction (*Heading 1*)

The Magnetic Tunnel Junction (MTJ) is one of the core building blocks for beyond-CMOS technologies. Compared to other beyond-CMOS building blocks, MTJs offer several unique advantages. Two factors stand out for their ability to enable computation-in-memory and other DoD applications. The first is the inherent non-volatility of the data storage layers, yielding essential radiation hardness, instant power-on, non-destructive data read-out, and security in hardware. And the second, is endurance up to 10^{16} read/write cycles and above, which is three orders of magnitude better than any reported memory cells. This provides necessary robustness for a true computation-in-memory and computation-near-memory arrays.

Advancement of MTJs in the past decade has led to successful magnetic random-access memory (MRAM) products, including toggle-MRAM (Everspin, IBM, Honeywell) and STT-MRAM (Everspin, GLOBALFOUNDRIES). It is realized by the community that MTJs may be the most promising building block for true computation in random access memory (CRAM) because of its superior endurance performance that is needed for computation [1-3]. State of the art performance metrics for the best performing p-MTJs have been demonstrated recently: ultrahigh TMR (208%) in p-MTJs with interfacial perpendicular magnetic anisotropy (PMA) and the ultralow write delays (t_{SW} ~ 165 ps at 50% switching probability) [4-5]. To enable true CRAM, as shown in Fig. 1, industry needs higher TMR ratio (>500%) and much lower write delay (<50 ps) with reasonably low operation energy (E_{SW} ~50 aJ per write operation, read is already negligibly low) and high thermal stability (delta = 60 k_BT). These numbers are from the preliminary benchmarking effort for CRAM. In theory, the highest TMR for a MgO-barrier based MTJ could reach

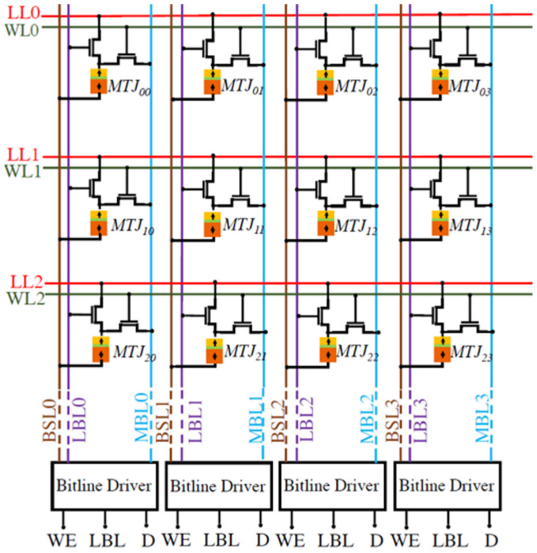


FIGURE 1: MRAM-based CRAM. LL, WL, BSL, LBL and MLM = Logic Line, Word Line, Bit Select Line, Logic Bit Line and Memory Bit Line

Gopman, Daniel; Wang, Jian-Ping; Low, Tony; Wang, Weigang; Taylor, William. "Toward Ultra-fast and Ultra-low Power Switching in Perpendicular Magnetic Tunnel Junctions." Paper presented at GomacTech 2020 Conference, San Diego, CA, US. March 16, 2020 - March 20, 2020.

35,000% with right materials and interfaces ,and the write delay could be reduced to below 10 ps with optimized materials and switching mechanisms. Our approach seeks to advance MTJs toward these performance specifications so that CRAM topology, many other new computing topologies, and future high-density MRAM products can become a reality.

An MTJ with engineered perpendicular magnetic anisotropy (PMA) relative to the planes of the layered heterostructure, or pMTJ, is an essential building block for advanced MTJs. The PMA materials are promising candidates for the development of ultra-high-density spintronic memory and logic devices due to their high thermal stability, scalability, and ultra-low power consumption. Interfacial PMA materials such as the Ta/CoFeB/MgO stack possess a PMA (K_u) of ~2-5 Merg/cm^3 and a α of ~0.015-0.027 [6-8]. Considerable progress has been made in the application of interfacial PMA materials to STT-MRAM [9-11]. Bulk PMA materials such as L1$_0$-FePd [12-16] and Mn-based perpendicular Heusler alloys [17-20] have a large K_u (1.3-1.4 MJ/m^3), a low α (0.002~0.015), and a low processing temperature (200 ºC). These properties make them ideal for ultra-high density and ultra-low power consumption memory devices. Furthermore, the write current density (J_c) for perpendicular spintronic devices is defined by the equation $J_c = 2\alpha e t_F M_s \left(H_{appl} + H_k\right) / \hbar\eta$ [21], where J_c mainly relates to the α, the saturation magnetization (M_S), and the K_u. Notably, switching time is proportional to M_S. However, for interfacial PMAs (such as the Ta/CoFeB/MgO stack) and bulk PMA FePd thin films, their M_S=1100~1300 kA/m, which is relatively high.

Synthetic antiferromagnetic (SAF) structures, comprised of two ferromagnetic layers aligned in an antiparallel arrangement and separated by a thin (<1 nm) metallic spacer, are used extensively in multilayered magnetic sensors and memories [22], which is one of promising methods to obtain the low M_s of the ferromagnetic layer for p-MTJs [23-25]. Owing to the nearly net-zero flux configuration, SAFs are generally implemented as the reference layer to compensate the deleterious offset fields from other layers. More recently, SAFs were demonstrated to show the potential for faster write operation in magnetic domain-wall memories compared with single-layer counterparts due to additional torques on the magnetization induced by the interlayer coupling field [26]. A recent simulation has made a similar argument for single-domain MRAMs in which a SAF *free layer* replaces the single magnetic layer [27]. This work, supported partially by the DARPA-sponsored STARnet research center C-SPIN, shows that the design of a SAF-MTJ using conventional interfacial PMA materials in construction of a SAF free layer can enable switching energy-delay products on the order of 100 aJ-10 ps for a 10 nm diameter free layer with a 60 k_BT thermal stability factor, where k_B is the Boltzmann constant and T is the ambient temperature (300 K). In addition, PI Jian-Ping Wang recently developed a perpendicular interfacial synthetic-ferrimagnetic structure (CoFeB/Gd/CoFeB) and realized field-free switching using spin orbit torques [28], and the bulk L1$_0$-FePd p-SAF composite free layer as well as the integration into p-MTJs and the 25% TMR ratio has been obtained at room temperature after 350 ºC post-annealing, even up to ~13% after 400 ºC post-annealing.

Ultrafast STT switching in the sub-ns regime is one of the key issues for STT-RAM development. One of the crucial limitations for ultrafast switching is the incubation delay induced by pre-switching oscillation [29]. Several approaches have been proposed to minimize pre-switching oscillations in order to improve the switching speed in spin valves such as developing all perpendicular structures [30], applying a hard axis field to set the free layer equilibrium away from the easy axis [31], and adding an extra perpendicular polarizer [32-34]. As of now, limited work has been done on sub-nanosecond STT switching in MTJs. Minimum switching times of 400–580 ps at 50% switching probability have been reported in conventional in-plane MTJs [35,36]. By adding a perpendicular polarizer, Liu et al showed 100% switching at 500 ps with external field assistance in their MTJ device [37]. Rowlands *et al* achieved 50% switching probability at 120 ps under zero bias field in a fully orthogonal MTJ [38]. For this present approach, a notable achievement is the observation of group of 165(190) ps at 50(98)% switching probabilities in the in-plane MTJs [5]. However, switching probabilities have not been studied extensively p-MTJs – with the exception being a single simulation work which showed that a SAF-MTJ using conventional interfacial PMA materials in construction of a SAF free layer can enable switching energy-delay products on the order of 100 aJ-10 ps for a 10 nm diameter free layer. With voltage controlled magnetic anisotropy (VCMA) [39,8], the switching energy of conventional pMTJs can be further reduced to sub-10 fJ level [40,41]. However, due to the precessional nature of the switching, the time of the pulse voltage has to be accurately controlled, giving rise to a write error rate (WER) of 10^{-5}, which far exceeds the tolerable limits for reliable chip-level performance [42]. We are exploring a unique Even-VCMA effect where voltages with both polarities can reduce the PMA; this is theoretically predicted when the Fermi level of the FM is precisely controlled [43]. With the advanced SAF structure developed in the proposal, STT and Even-VCMA can work together in a complimentary fashion, eventually leading to ultra-low energy (<100 aJ), ultra-low WER (< 10^{-10}) and ultra-fast (< 50 ps) switching.

The key specifications for advanced pMTJs for computational random access memory systems are: ultra-high TMR ratio (>500%); ultra-fast switching (50 ps for 50% switching probability of a 60 k_BT pMTJ); ultra-low switching energy (100 aJ); ultra-small feature size (10 nm) and low damping constant for the magnetic storage layer (<0.01). This provides necessary robustness for a true computation-in-memory and computation-near-memory arrays.

With the realization of advanced pMTJs for computational random access memory systems, significant gains can be realized over classic, near- and edge memory computing. By using memory cells to carry out computation, one gains a 1400-fold improvement over near-memory computing in execution time, and a 40-fold energy reduction [44]. Surprisingly, our team has discovered a realization of a spintronic, reconfigurable in-memory binary neural network accelerator that performs with greater energy efficiency and throughput on image classification and genomics kernel tasks than corresponding CPU-, GPU- and FPGA-based implementations [45].

II. Achieving ultrafast switching MTJs

To reach ultimate fast switching, a balanced SAF (with near zero net magnetization) is desired. However, such as balanced SAF would require an extremely large exchange coupling constant which is not practical. Therefore, aside from the gains in write speed realized by the SAF structure, additional reduction in write delay may be achieved by electric-field tuning of the PMA using the so-called even voltage control of magnetic anisotropy (even-VCMA) effect.

Indeed, a promising approach to reduce the overall switching energy (E_{SW}) in MTJs is utilizing the interfacial perpendicular magnetic anisotropy (PMA) at the FM/oxide interface and the corresponding VCMA effect [8,36-38,46-48], where the MTJ can be precessionally switched by an effective in-plane field generated by a sub-ns voltage pulse. We have achieved a low switching current at 10^8-10^9 A/m^2 with fast speed at 400 ps, corresponding to a low E_{SW} that is below 20 fJ. However, due to the precessional nature of switching, the MTJ can be only be switched in a very tight window in VCMA, giving rising to a very large write error rate in the order of 10^{-5}, much larger than STT (10^{-10}). The core structure is CoFeB (0.9 nm)/MgO (1.6 nm)/CoFeB (1.5 nm), with thicknesses reported in parentheses. However, in another device with a very similar structure, we were able to extend the switching window from 0.5 ns to more than 1 ns with external field an applied in-plane instead of at 40°.

The key to reducing switching energy and switching time with interfacial SAF free layer is to utilize the VCMA effect to lower the energy barrier during switching. The efficiency of VCMA is characterized by the change of interfacial magnetic anisotropy density per unit electric field, $\Delta K_S/\Delta E_{ext}$. With the Even-VCMA effect where the STT and VCMA can be combined, truly ultra-low energy and ultra-fast switching is possible, simultaneously with a very low WER that is comparable to STT alone.

III. Development of SAF MTJs

A. Interfacial PMA MTJ

Consistent with current state-of-the art approaches, we have designed p-MTJs with interfacial perpendicular SAF free layers for sub-100 ps switching (see Fig. 2(a)). We have already achieved above 200% TMR with Mo dusting layers inserted at the Ta/CoFeB interface [4]. Unlike thick Mo layers that exhibited a strong (110) crystalline texture, the inserted Mo layer between Ta/CoFeB had little negative influence on the crystallization of CoFe (001), thereby combining the advantages of Mo as a good thermal barrier and Ta as a good boron sink. We envision possible solutions using other heavy metal insertions, to include W and Hf, which have potential to provide higher TMR.

The high probability switching of interfacial p-MTJs has been demonstrated in 100 nm diameter nanopillars. Our p-MTJs comprising a Ta/CoFeB/MgO free layer, rapid thermal annealed for 5 minutes at 300 °C exhibit a significant VCMA effect that enables fast, reliable switching at 200 ps. We report switching probability greater than 90% for a 200 ps pulse and 40 fJ switching energy. By engineering a thicker MgO tunneling barrier (approximately 50% thicker than the previous sample), the switching energy of a 100 nm p-MTJ can be reduced to 9 fJ, with a modest reduction in switching probability.

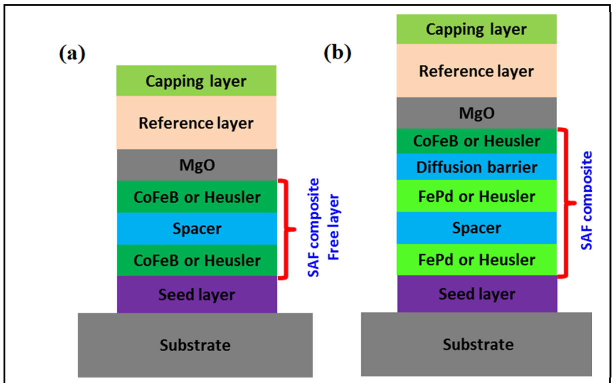


FIGURE 2: (a) The designed interfacial perpendicular SAF(p-SAF) structure, in which the CoFeB and Co-based in-plane Heusler alloys will be used as a ferromagnetic layer and Ru, Ta, Cr will be as a spacer to generate the synthetic-AFM, and Gd will be a spacer to lead to synthetic-ferrimagnetic layer. (b) The designed bulk p-SAF structure with L1$_0$-FePd and Mn-based perpendicular Heusler alloys as a ferromagnetic layer and Ru and Cr as a spacer.

B. Bulk PMA MTJ

The p-SAF structure with bulk PMA FePd thin films has been developed and integrated into p-MTJs as a free layer (see Fig. 2(b)). The p-SAF structure using bulk PMA FePd thin films through an fcc phase Ru spacer were designed and developed. It can be found that the L1$_0$-phase FePd p-SAF structure presents good PMA and antiferromagnetic coupling properties with a square shape minor M-H loop and a net remanent magnetization ~500 kA/m. The PMA constant K_u of the FePd p-SAF structure was then evaluated to be ~1.05 MJ/m^3 based on its M_s and H_K from the M-H loops, which is several times larger than that of interfacial PMA materials. The J_{iec} of the FePd p-SAF structure was calculated to be ~ -2.86 mJ/m^2, which is about one order of magnitude larger than that of the [Co/Pd]$_n$ p-SAF system with the same post-annealing temperature [50]. In addition, using magnified STEM we demonstrated that the Ru spacer follows the texture of FePd layer and forms the metastable fcc phase. In addition, most recently the p-MTJs with the [Co/Pt]$_n$ SAF layer from Japanese group showing high TMR (131% by CIPT after 350 °C post-annealing)[51], however, compared with FePd (α~0.002) [11], the [Co/Pt]$_n$ multilayer has a much larger α ~0.05~0.18 which will result in an appreciably larger switching current density which translates to higher write energy and slower write times.

Gopman, Daniel; Wang, Jian-Ping; Low, Tony; Wang, Weigang; Taylor, William. "Toward Ultra-fast and Ultra-low Power Switching in Perpendicular Magnetic Tunnel Junctions." Paper presented at GomacTech 2020 Conference, San Diego, CA, US. March 16, 2020 - March 20, 2020.

IV. Benchmarking and metrics of SAF MTJs for fast switching

The relevant metrics of SAF pMTJs for fast switching and implementation as advanced MTJs for computation in random access memory have informed the development activities described above. These include device size, switching delay, switching energy, write error rate, thermal stability, tunneling magnetoresistance and the damping constant. For the interfacial p-MTJs, these results are specific to the CoFeB/MgO/CoFeB p-MTJ system. Here we have successfully engineered 100 nm nanopillars with energy-delay product 9 fJ-200 ps, and that are thermally stable with TMR exceeding 100%. The switching probability of these devices was 80% and damping constant 0.01.

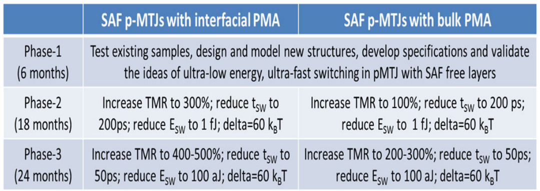
For the bulk SAF pMTJs, strategies have been developed to identify the useful buffer layers that can engineer a large perpendicular magnetic anisotropy in FePd, thereby providing thermal stability down to 10 nm diameter nanopillars. We find that these include Pt, Ru, Ir, Rh, but not Mo or Ta. This is consistent with the need for a fcc (001) buffer whose in-plane lattice parameter is comparable with the in-plane lattice parameter of FePd (0.385 nm). We have also shown the ability to engineer a bulk SAF trilayer using FePd/Ru/FePd, FePd/Ir/FePd and FePd/Rh/FePd. We find that Ru and Ir spacers are more thermally stable over the high-temperature growth and annealing treatment than the Rh spacer and can convey a net zero magnetization state at zero applied field, a prerequisite of a SAF device.

V. Conclusions

To deliver on the potential of MRAM for computation-in-memory, reductions in the energy- and delay characteristics of pMTJs must be demonstrated. Based on interfacial- and bulk perpendicular magnetic anisotropy materials, we believe there is a path to developing two novel perpendicular synthetic antiferromagnet (p-SAF) designs for ultra-fast and ultra-low power switching performance: one using interfacial PMA materials and one using bulk PMA materials. Our stacks are compatible with or close to the existing p-MTJ stack and fabrication process, which can transition to back-end-of-line semiconductor process practical within a 5-10 year time frame within our near-term roadmap shown in Fig. 3.

Acknowledgment *(Heading 5)*

This work is funded in part by DARPA HR001117S0056—FP—042 "Advanced MTJs for computation in and near random access memory," and by NIST.



	SAF p-MTJs with interfacial PMA	SAF p-MTJs with bulk PMA
Phase-1 (6 months)	Test existing samples, design and model new structures, develop specifications and validate the ideas of ultra-low energy, ultra-fast switching in pMTJ with SAF free layers	
Phase-2 (18 months)	Increase TMR to 300%; reduce t_{SW} to 200ps; reduce E_{SW} to 1 fJ; delta=60 k_BT	Increase TMR to 100%; reduce t_{SW} to 200 ps; reduce E_{SW} to 1 fJ; delta=60 k_BT
Phase-3 (24 months)	Increase TMR to 400-500%; reduce t_{SW} to 50ps; reduce E_{SW} to 100 aJ; delta=60 k_BT	Increase TMR to 200-300%; reduce t_{SW} to 50ps; reduce E_{SW} to 100 aJ; delta=60 k_BT

FIGURE 3: Roadmap to demonstration of advanced MTJs for CRAM.

References

[1] H. Meng, J. Wang and J. P. Wang "A Spintronics Full Adder for Magnetic CPU", IEEE Electron Dev. Lett., 26, 360 (2005).

[2] A. Lyle, S. Patil, et al, "Direct Communication Between Magnetic Tunnel Junctions for Non-Volatile Logic Fan-Out Architecture," Appl. Phys. Lett. 97, 152504 (2010)

[3] J. P. Wang and J. D. Harms, "General structure for computational random access memory (CRAM)," 2013. US Patent 9,224,447 B2.

[4] H. Almasi, M. Xu, Y. Xu, T. Newhouse-Illige, and W. Wang, "Effect of Mo insertion layers on the magnetoresistance and perpendicular magnetic anisotropy in Ta/CoFeB/MgO junctions." Appl. Phys. Lett., 109(3), 032401 (2016).

[5] H. Zhao, A. Lyle, Y. Zhang, P. K. Amiri, G. Rowlands, Z. Zeng, J. Katine, H. Jiang, K. Galatsis, K. L. Wang, I. N. Krivorotov, and J.-P. Wang, "Low writing energy and sub nanosecond spin torque transfer switching of in-plane magnetic tunnel junction for spin torque transfer random access memory," J. Appl. Phys. 109, 07C720 (2011)

[6] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura and H. Ohno, "A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction," Nat. Mater. 9, 721 (2010).

[7] H. Sato, M. Yamanouchi, K. Miura, S. Ikeda, H. D. Gan, K. Mizunuma, R. Koizumi, F. Matsukura, and H. Ohno, "Junction size effect on switching current and thermal stability in CoFeB/MgO perpendicular magnetic tunnel junctions," Appl. Phys. Lett. 99, 042501 (2011).

[8] S. Iihama, S. Mizukami, H. Naganuma, M. Oogane, Y. Ando, and T. Miyazaki, "Gilbert damping constants of Ta/CoFeB/MgO(Ta) thin films measured by optical detection of precessional magnetization dynamics," Phys. Rev. B 89, 174416 (2014).

[9] P. Khalili Amiri, Z. M. Zeng, J. Langer, H. Zhao, G. Rowlands, Y.-J. Chen, I. N. Krivorotov, J.-P. Wang, H. W. Jiang, J. A. Katine, Y. Huai, K. Galatsis, and K. L. Wang, "Switching current reduction using perpendicular anisotropy in CoFeB-MgO magnetic tunnel junctions," Appl. Phys. Lett. 98, 112507 (2011).

[10] H. Sato, E. C. I. Enobio, M. Yamanouchi, S. Ikeda, S. Fukami, S. Kanai, F. Matsukura, and H. Ohno, "Properties of magnetic tunnel junctions with a MgO/CoFeB/Ta/CoFeB/MgO recording structure down to junction diameter of 11 nm," Appl. Phys. Lett. 105, 062403 (2014).

[11] W.-G. Wang, M. Li, S. Hageman, and C. L. Chien, "Electric-field-assisted switching in magnetic tunnel junctions," Nat. Mater. 11, 64 (2012).

[12] D. Weller, A. Moser, L. Folks, M. E. Best, W. Lee, M. F. Toney, M. Schwickert, and J.-U. Thiele, M. F. Doerner, "High Ku Materials Approach to 100 Gbits/in^2," IEEE Trans. Magn. 36, 10 (2000).

[13] S. Iihama, A. Sakuma, H. Naganuma, M. Oogane, T. Miyazaki, S. Mizukami, and Y. Ando, "Low precessional damping observed for L1$_0$-ordered FePd epitaxial thin films with large perpendicular magnetic anisotropy," Appl. Phys. Lett. 105, 142403 (2014).

[14] S. Iihama, A. Sakuma, H. Naganuma, M. Oogane, S. Mizukami, and Y. Ando, "Influence of L1$_0$ order parameter on Gilbert damping constants for FePd thin films investigated by means of time-resolved magneto-optical Kerr effect," Phys. Rev. B 94, 174425 (2016).

[15] H. Naganuma, G. Kim, Y. Kawada, N. Inami, K. Hatakeyama, S. Iihama, K. M. N. Islam, M. Oogane, S. Mizukami, and Y. Ando, "Electrical detection of millimeter-waves by magnetic tunnel junctions using perpendicular magnetized L1$_0$-FePd free layer," Nano Lett. 15, 623 (2015).

[16] S. Iihama, M. Khan, H. Naganuma, M. Oogane, T. Miyazaki, S. Mizukami, and Y. Ando, "Magnetization Dynamics and Damping for L1$_0$-FePd Thin Films with Perpendicular Magnetic Anisotropy," J. Magn. Soc. Jpn. 39, 57 (2015).

[17] H. Kurt, K. Rode, M. Venkatesan, P. Stamenov, and J. M. D. Coey, "High spin polarization in epitaxial films of ferrimagnetic Mn$_3$Ga," Phys. Rev. B 83, 020405(R) (2011).

[18] S Mizukami, F Wu, A Sakuma, J Walowski, D Watanabe, T Kubota, X Zhang, H Naganuma, M Oogane, Y Ando, T Miyazaki, "Long-lived ultrafast spin precession in manganese alloys films with a large perpendicular magnetic anisotropy," Phys. Rev. Lett. 106, 117201(2011).

[19] S. Mizukami, A. Sugihara, S. Iihama, Y. Sasaki, K. Z. Suzuki, T. Miyazaki, “Laser-induced THz magnetization precession for a tetragonal Heusler-like nearly compensated ferromagnet,” Appl. Phys. Lett. 108, 012404 (2016).

[20] T. Kubota, Q. L. Ma, S. Mizukami, X. M. Zhang, H. Naganuma, M. Oogane, Y. Ando and T. Miyazaki, “Magnetic tunnel junctions of perpendicularly magnetized L1$_0$-MnGa/Fe/MgO/CoFe structures: Fe-layer-thickness dependences of magnetoresistance effect and tunnelling conductance spectra,” J. Phys. D: Appl. Phys. 46 155001(2013).

[21] Z. Diao, Z. Li, S. Wang, Y. Ding, A.Panchula, E. Chen, L.-C. Wang and Y. Huai, “Spin-transfer torque switching in magnetic tunnel junctions and spin- transfer torque random access memory,” J. Phys.: Condens. Matter 19, 165209 (2007).

[22] R. A. L. Duine, K.-J. Lee, S. S. P. Parkin, and M. D. Stiles, “Synthetic Antiferromagnetic Spintronics,” arXiv:1705.10526, 2017

[23] A. Bergman, B. Skubic, J. Hellsvik, L. Nordström, A. Delin, and O. Eriksson, “Ultrafast switching in a synthetic antiferromagnetic magnetic random-access memory device,” Phys. Rev. B 83, 224429 (2011).

[24] C. Y. You, “Effect of the synthetic antiferromagnetic polarizer layer rigidness on the spin transfer torque switching current density,” Appl. Phys. Lett. 103, 042402 (2013).

[25] X. G. Xu, D. L. Zhang, X. Q. Li, J. Bao, Y. Jiang, and M. B. A. Jalil, “Synthetic antiferromagnet with Heusler alloy ferromagnetic layers,” J. Appl. Phys. 106, 123902 (2009).

[26] S. H. Yang, K. S. Ryu, and S. Parkin, “Domain-wall velocities of up to 750 m s-1 driven by exchange-coupling torque in synthetic antiferromagnets,” Nature Nanotechnology, 10, 221 (2015).

[27] K. Y. Camsari, A. Z. Pervaiz, R. Faria, E. E. Marinero, and S. Datta, “Ultrafast Spin-Transfer-Torque Switching of Synthetic Ferrimagnets,” IEEE Magn. Lett. 7, 3107205 (2016).

[28] J.-Y. Chen, Mahendra DC, D. L. Zhang, Z. Zhao, M. Li, and J.-P. Wang, “Field-free spin-orbit torque switching of composite perpendicular CoFeB/Gd/CoFeB layers utilized for three-terminal magnetic tunnel junctions,” Appl. Phys. Lett. 111, 012402 (2017)

[29] T. Devolder, C. Chappert, J. A. Katine, M. J. Carey, and K. Ito, “Distribution of the magnetization reversal duration in subnanosecond spin-transfer switching,” Phys. Rev. B 75, 064402 (2007)

[30] D. Bedau, H. Liu, J.-J. Bouzaglou, A. D. Kent, J. Z. Sun, J. A. Katine, E. E. Fullerton, and S. Mangin, “Ultrafast spin-transfer switching in spin valve nanopillars with perpendicular anisotropy,” Appl. Phys. Lett. 96, 022514 (2010)

[31] T. Devolder, P. Crozat, J.-V. Kim, and C. Chappert, K. Ito, J. A. Katine and M. J. Carey, “Magnetization switching by spin torque using subnanosecond current pulses assisted by hard axis magnetic fields,” Appl. Phys. Lett. 88, 152502 (2006)

[32] C. Papusoi, B. Delaët, B. Rodmacq, D. Houssameddine, J.-P. Michel, U. Ebels, R. C. Sousa, L. Buda-Prejbeanu, and B. Dieny, “100 ps precessional spin-transfer switching of a planar magnetic random access memory cell with perpendicular spin polarizer,” Appl. Phys. Lett. 95, 072506 (2009)

[33] O. J. Lee, V. S. Pribiag, P. M. Braganca, P. G. Gowtham, D. C. Ralph, and R. A. Buhrman, “Ultrafast switching of a nanomagnet by a combined out-of-plane and in-plane polarized spin current pulse,” Appl. Phys. Lett. 95, 012506 (2009);

[34] J-M. L. Beaujour, D. B. Bedau, H. Liu, M. R. Rogosky, A. D. Kena, “Spin-transfer in nanopillars with a perpendicularly magnetized spin polarizer,” Proc. SPIE 7398 73980D (2009).

[35] T. Aoki, Y. Ando, M. Oogane, and H. Naganuma, “Reproducible trajectory on subnanosecond spin-torque magnetization switching under a zero-bias field for MgO-based ferromagnetic tunnel junctions,” Appl. Phys. Lett. 96, 142502 (2010)

[36] H. Liu, D. Bedau, D. Backes, J. A. Katine, J. Langer, and A. D. Kent, “Ultrafast switching in magnetic tunnel junction based orthogonal spin transfer devices,” Appl. Phys. Lett. 97, 242510 (2010)

[37] G. E. Rowlands, T. Rahman, J. A. Katine, J. Langer, A. Lyle, H. Zhao, J. G. Alzate, A. A. Kovalev, Y. Tserkovnyak, Z. M. Zeng, H. W. Jiang, K. Galatsis, Y. M. Huai, P. Khalili Amiri, K. L. Wang, I. N. Krivorotov, and J.-P. Wang, “Deep subnanosecond spin torque switching in magnetic tunnel junctions with combined in-plane and perpendicular polarizers,” Appl. Phys. Lett. 98, 102509 (2011)

[38] F. Matsukura, Y. Tokura, and H. Ohno, “Control of magnetism by electric fields,” Nat. Nanotechnol. 10, 209, (2015).

[39] C. Grezes, F. Ebrahimi, J. G. Alzate, X. Cai, J. A. Katine, J. Langer, B. Ocker, P. Khalili, and K. L. Wang, “Ultra-low switching energy and scaling in electric-field-controlled nanoscale magnetic tunnel junctions with high resistance-area product,” Appl. Phys. Lett. 108, 12403 (2016).

[40] S. Kanai, F. Matsukura, and H. Ohno, “Electric-field-induced magnetization switching in CoFeB/MgO magnetic tunnel junctions with high junction resistance,” Appl. Phys. Lett. 108, 192406 (2016).

[41] J. J. Nowak, R. P. Robertazzi, J. Z. Sun, G. Hu, J. H. Park, J. Lee, A. J. Annunziata, G. P. Lauer, R. Kothandaraman, E. J. O Sullivan, P. L. Trouilloud, Y. Kim, and D. C. Worledge, “Dependence of Voltage and Size on Write Error Rates in Spin-Transfer Torque Magnetic Random-Access Memory,” IEEE Magn. Lett. 7, 4 (2016).

[42] W. Skowroński, T. Nozaki, D. D. Lam, Y. Shiota, K. Yakushiji, H. Kubota, A. Fukushima, S. Yuasa, and Y. Suzuki, “Underlayer material influence on electric-field controlled perpendicular magnetic anisotropy in CoFeB/MgO magnetic tunnel junctions,” Phys. Rev. B 91, 184410 (2015).

[43] M. Weisheit, S. Fähler, A. Marty, Y. Souche, C. Poinsignon, and D. Givord, “Electric field-induced modification of magnetism in thin-film ferromagnets,” Science **315**, 349 (2007).

[44] Y. Shiota, T. Nozaki, F. Bonell, S. Murakami, T. Shinjo, and Y. Suzuki, “Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses,” Nat. Mater. **11**, 39 (2012).

[45] M. Zabihi, Z. Chowdhury, Z. Zhao, U. R. Karpuzcu, J. Wang, S. Sapatnekar, "In-memory processing on the spintronic CRAM: From hardware design to application mapping", IEEE Trans. Comput., 68, 1159 (2018)

[46] D. L. Zhang, C. Sun, Y. Lv, K. B. Schliep, Z. Zhao, J.-Y. Chen, P. M. Voyles, and J.-P. Wang, “L1$_0$-FePd Synthetic Antiferromagnet Through a Face-centered-cubic Ruthenium Spacer Utilized for Perpendicular Magnetic Tunnel Junctions,” 9, 044028 (2018).

[47] Salonik Resch, S. Karen Khatamifard, Zamshed Iqbal Chowdhury, Masoud Zabihi, Zhengyang Zhao, Jian-Ping Wang, Sachin S. Sapatnekar, and Ulya R. Karpuzcu. 2019. PIMBALL: Binary Neural Networks in Spintronic Memory. ACM Trans. Archit. Code Optim. 16, 4, 41 (2019).

[48] J.-B. Lee, G.-G. An, S.-M. Yang, H.-S. Park, W.-S. Chung, and J.-P. Hong, “Thermally robust perpendicular Co/Pd-based synthetic antiferromagnetic coupling enabled by a W capping or buffer layer,” Scientific Reports 6, 21324 (2016).

[49] D.-L. Zhang, M. Bapna, W. Jiang, D. P. de Sousa, C. Y. Liao, Z. Zhao, Y. Lv, A. Naeemi, T. Low, S. A. Majetich, J.-P. Wang. Bipolar electric-field switching of perpendicular magnetic tunnel junctions through voltage-controlled exchange coupling. arXiv preprint arXiv:1912.10289. 2019 Dec 21

[50] K. Yakushiji, A. Sugihara, A. Fukushima, H. Kubota, and S. Yuasa, “Very strong antiferromagnetic interlayer exchange coupling with iridium spacer layer for perpendicular magnetic tunnel junctions,” Appl. Phys. Lett. 110, 092406 (2017)

[51] A. Caprile, M. Pasquale, M. Kuepferling, M. Co¨ısson, T. Y. Lee, and S. H. Lim, “Microwave Properties and Damping in [Pt/Co] Multilayers With Perpendicular Anisotropy,” IEEE Magnetic Letters, 5, 3000304(2014)

Proceedings of the ASME 2020 International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
IDETC/CIE 2020
August 16-19, 2020, St. Louis, MO, USA

IDETC2020-22300

LINKING PERFORMANCE DATA AND GEOSPATIAL INFORMATION OF MANUFACTURING ASSETS THROUGH STANDARD REPRESENTATIONS

Aaron Hanke[†,♣], **Teodor Vernica**[†,♠], **William Z. Bernstein**[‡*]
[†]Associate, Systems Integration Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899
[♣]TU-Dresden, Chair of Engineering Design and CAD, 01069 Dresden, Germany
[♠]Aarhus University, Department of Computer Science, 8200 Aarhus N
[‡]Systems Integration Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899
Email: {aaron.hanke, teodor.vernica, william.bernstein}@nist.gov

ABSTRACT

Interoperability across emerging visualization modalities, including augmented reality (AR) and virtual reality (VR), remains a challenge with respect to industrial applications. One critical issue relates to the lack of standard approaches for coordinating geospatial representations that are required to facilitate AR/VR scenes with domain-specific information in the form of time-series data, solid models, among other data types. In this paper, we focus on the linking of manufacturing asset data via the MTConnect standard with geospatial data via the IndoorGML standard. To this end, we demonstrate the utility of this integration through two visualization-based prototype implementations, including one focused on (i) monitoring production facilities to improve situational awareness and (ii) evaluating and delivering suggested navigation paths in production facilities. We then comment on implications of such standards-driven approaches for related domains, including AR prototype development and automatic guided vehicles.

Keywords: Geospatial modeling, standards, augmented reality, smart manufacturing

1 Introduction

The digitalization of manufacturing systems has become a core focus area to improve agility, productivity, and efficiency. This trend is especially important for introducing emerging information technologies (IT), e.g., virtual reality (VR) and augmented reality (AR), into production scenarios. The combination of advanced manufacturing with IT capabilities is referred to as *smart manufacturing* [1]. Early adopters of smart manufacturing concepts have been able to lower costs, improve worker safety, and achieve deeper situational awareness [2]. To gain broad use and dissemination of such solutions, standards are vital since the collaborative communication and integration of the underlying hardware and software is critical for success [3]. However, interoperability of such applications, methods, and devices remains challenging for the broad realization of smart manufacturing [4].

In this paper, we aim to address one of the interoperability problems faced by manufacturing organizations looking to adopt technologically advanced solutions. As an anchor to more seamless integration, we focus on the representation (i.e., computer-readable data model) of geospatial definitions. Linking real-time geospatial data across production systems is critical for the proper coordination of (i) automatic guided vehicles (AGVs) [5], (ii) interface-guided navigation for human operators [6], and (iii) perspectives for higher level situational awareness for decision makers [7]. Considering the ubiquity of on-board spatial reasoning devices, e.g., Light Detection and Ranging (LiDAR) technologies on AGVs, aligning multiple coordinate systems, facilitating multiple user perspectives, and tracking multiple moving devices is not trivial.

In response, we propose the use of standards for static

[*] Address all correspondence to this author.

Hanke, Aaron; Vernica, Teodor; Bernstein, William Z. "Linking Performance Data and Geospatial Information of Manufacturing Assets through Standard Representations." Paper presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

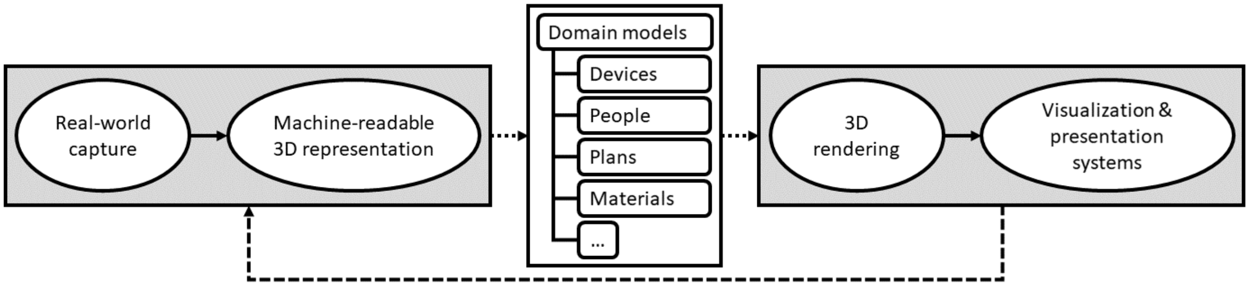


Figure 1. General pipeline for generating industrial visualization prototypes. Activities in grey and white boxes summarize standards efforts from the visualization and SMS communities, respectively. Dashed lines signal opportunities for automation.

geospatial representations that can build a bridge between such vision-based systems. As a first step, in this paper, we study the feasibility of leveraging one such standard, i.e., IndoorGML[1] released by the Open Geospatial Consortium (OGC), with an existing Smart Manufacturing Systems (SMS) standard, i.e., MTConnect[2], to facilitate the geospatial underpinnings of indoor manufacturing environments. We chose to work with these specific standards due to the publicly available tools supporting IndoorGML as well as publicly available MTConnect data via the NIST SMS Testbed (see Sec. 4).

Through this demonstration, we showcase two use cases derived from the underlying mapping between the standards. The first use case shows the quick three-dimensional (3D) rendering and presentation of a shop floor for situational awareness. The second use case presented demonstrates a tablet-based guidance for navigation from one job to another on the shop floor. Both use cases were generated by leveraging existing tools and methods already produced in accordance with the standards. In our minds, this demonstrates the power of standards in that leveraging tools that interface with standard data representation is much easier than building each component from scratch. To conclude the paper, we discuss implications of building additional bridges, i.e., mappings and toolkits, across standard interfaces to improve VR, AR, and mixed reality (XR) interoperability.

2 Motivation and Background

Industrial visualization-driven installations, e.g., AR and XR implementations, are often the production of one-off prototypes, wherein the domain models, e.g., machining performance models, digital solid models, and user manuals, are tightly coupled with domain-agnostic interfaces, e.g., rendering modules, presentation modalities, and visualization engines. After years of many organizations developing their own one-off installations, interoperability has become more of a pipe-dream. Solutions to help solve the problem are severely lacking. To address such needs, standards development organizations, such as IEEE, OGC, and the Khronos Group, have worked to contribute standard representations, modules, and languages. However, these efforts suffer from severe silo-ing and seem to fail to communicate with one another. This is especially true for domain-agnostic groups, e.g., World Wide Consortium (W3C) and Khronos Group, communicating with domain-heavy groups, e.g., American Society of Mechanical Engineers (ASME), the MTConnect Institute, and the OPC Foundation.

However, both perspectives, domain-specific thinking, e.g., for manufacturing and field maintenance, and visualization-specific concerns, e.g., real-world capture and scene rendering, are vital. SMS-specific standards, e.g., MTConnect and OPC-UA, provide the necessary semantic descriptions of concepts, such as information about devices, people, and materials. Figure 1 showcases the current state of industrial prototype development. From a high-level view, the visualization community is focused on two separate efforts, (1) digitizing real-world information (shown in the box to the left of Fig. 1) and (2) rendering and presenting scenes through the appropriate visualization modalities (see box to the right of Fig. 1). To produce successful and meaningful user experiences, it is vital to connect to domain-specific models. However, in the current state, automating these data transformations is expert-driven and requires many iterations and human hours. To this end, there is significant opportunities in automating (or self-automating) these transformations (indicated as dashed lines across Fig. 1).

In this paper, we particularly focus on the potential of standards to help contextualize geospatial information with streaming data collected form the shop floor. Below, we review relevant work related to (1) representing indoor spaces specifically for production environments and (2) leveraging those representations for navigation purposes.

[1] Available at https://www.indoorgml.net/

[2] Available at https://www.mtconnect.org/

2.1 Indoor Space Definition for SMS

Since the majority of manufacturing processes are located inside buildings [8], modeling indoor spaces is required to fully represent production systems virtually. Often, the initial description of such environments begins with a map or layout of the workshop floor. Such maps are helpful in (1) supporting AGV tasks, (2) tracking tools, components and fixtures, as well as (3) facilitating guided navigation for operators [9, 10].

One of the core challenges of leveraging maps of production systems is the disparate nature of their source and intent. For example, AGVs, depending on brand and purpose, construct their own maps at various levels of detail [11]. In such cases, some AGVs leverage on-board LiDAR systems to build maps in order to avoid walls and other obstacles. To curate such maps, additional manual labor is required. Other AGVs have used camera systems that read optical markers on the floor or the ceiling to follow a predefined path through the workshop [12, 13]. Both solutions work for their specific use case, but the created maps cannot be interchanged easily. This example demonstrates that even in cases of similar technologies, e.g., two different types of AGVs, relating geospatial data and aligning coordinate systems represents significant redundant work.

In complex environments that are ubiquitous in SMS, the overhead of reconstructing geospatial definitions is significant. As a result, there exist standard efforts for constructing and curating data representation that define such geospatial information. Here, we specifically focus on modeling indoor spaces. Standards have been used across domains to limit redundant work. The most widely used standards for representing indoor spaces relate to Building Information Model (BIM) activities, an approach with a suite of standards built using the Industry Foundation Classes (IFC) [14, 15]. The IFC standards and developed and maintained by buildingSMART[3]. To support complex BIM modeling, there are other standard data representations that facilitate lower level modeling. In this paper, we specifically focus on two OGC standards, namely CityGML[4] and IndoorGML[5]. More information about each standard is provided below in Sec. 3.

Defining the boundaries of indoor spaces is not complete without the full expression of affixed and mobile objects in that space. This is especially true in production environments, wherein the location of objects, such as handheld tools and devices, is a practical challenge [16, 17]. Carrasco et al. [18] argues that there is a strong need for low-cost solutions for indoor localization technologies to support small and medium enterprises (SMEs). Such technologies include WiFi or Bluetooth Low Energy (BLE)-based location mapping. The downside of these technologies is that they can only deliver an accuracy of about 1-2 meters. To facilitate more precise positioning, there are systems that utilize RFID chips, Ultra Wideband (UWB), Indoor GPS, or simultaneous localization and mapping (SLAM)-based approaches. These systems are in general more expensive, but can deliver accuracy down to a few centimeters.

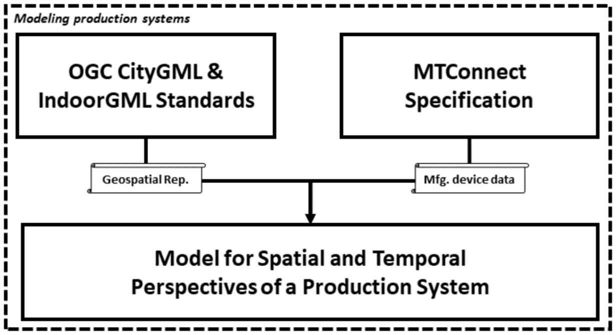

Modeling production systems
OGC CityGML & IndoorGML Standards
MTConnect Specification
Geospatial Rep.
Mfg. device data
Model for Spatial and Temporal Perspectives of a Production System

Figure 2. One instance of combining two disparate standards for quick AR prototype deployment for situational awareness and indoor navigation in smart manufacturing systems. We focus on leveraging OGC standards in concert with MTConnect.

2.2 Indoor Space Navigation for SMS

With a geospatial definition of an indoor space, an apparent use case of such a representation is centered around navigation purposes. Referencing a formal description of indoor building entities, such as walls, structural columns, corridors, and other obstacles, finding physically viable paths can be facilitated. Navigation tasks are commonly derived in SMS across a variety of clients. For example, to accomplish their intended tasks, such as material transport [19, 8], AGVs must navigate complex manufacturing environments and avoid unforeseen obstacles, e.g., plastic barriers, ramps, and other devices.

Since maintenance costs of buildings are much higher than its construction costs [20], balancing expenses of maintenance tasks is important. Lee et al. [21] observed and analysed tradespeople in maintenance fieldwork to investigate their current practices. They showed that up to 50% of time in maintenance is spent on field navigation and object localization. In response, they developed an AR-based software application [6] that utilizes an Operations and Maintenance (O&M) information model to support O&M fieldwork. Their prototype showed a reduction of up to 51% of time needed to locate target areas. Many more examples exist for how computer-supported, often AR-based [22], maintenance tasks require indoor navigation approaches.

3 IndoorGML – MTConnect Integration

Figure 3 depicts the main goal of this paper to achieve a standards-driven model for spatial and temporal perspectives of

[3] Available at https://technical.buildingsmart.org/standards/bcf/
[4] Available at https://www.opengeospatial.org/standards/citygml
[5] Available at https://www.opengeospatial.org/standards/indoorgml

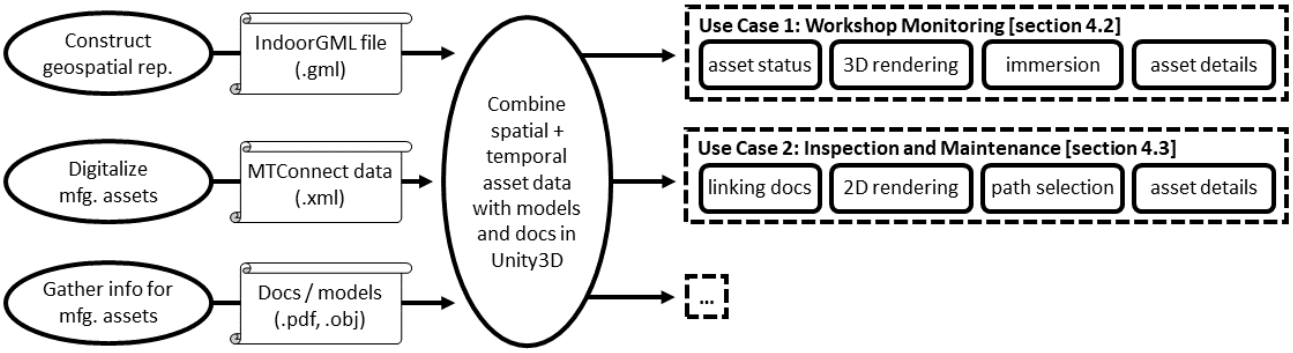

Construct geospatial rep.
IndoorGML file (.gml)
Digitalize mfg. assets
MTConnect data (.xml)
Gather info for mfg. assets
Docs / models (.pdf, .obj)
Combine spatial + temporal asset data with models and docs in Unity3D
Use Case 1: Workshop Monitoring [section 4.2]
asset status
3D rendering
immersion
asset details
Use Case 2: Inspection and Maintenance [section 4.3]
linking docs
2D rendering
path selection
asset details
...

Figure 3. Process workflow of our implementation. In parallel, we construct a geospatial representation of the SMS Test Bed, access near-real-time data from the MTConnect data streams, and obtain auxiliary information about the connected devices. Through Unity3D, we achieve the IndoorGML-MTConnect integration to facilitate the discussed use cases.

a production system. In this section, we describe each standard's data model and considerations for their integration.

IndoorGML is a standard developed by OGC that describes indoor spatial information in an open data model within an eXtensible Modeling Language (XML) Schema Definition (XSD) [15]. The standard focuses on modeling indoor spaces including its properties and the underlying topology. Intentionally, this standard was designed to be used for indoor location-based services, such as navigation or geo-tagging. One of the core concepts of IndoorGML is the definition of rooms as cells with its properties, e.g., name, ID, and purpose of the room, as well as whether it is navigable or non-navigable. Cells are not allowed to overlap each other, but they can share a common boundary. Additionally, it is possible to define nodes (or states) which can be connected by edges (or transitions). They are mostly used for indoor navigation where the transitions operate as paths and the states as way-points. IndoorGML provides the possibility to define topological information of cells by creating a Node-Relation-Graph (NRG). The NRG describes the connectivity and adjacency of cells and simplifies the 3D representation of an indoor space to enable use by complex computational processes. To create a NRG, the Poincaré duality is used. With that a k-dimensional object can be transferred into an n-dimensional space by creating an (n-k)-dimensional object. For example a 3D cell becomes a zero-dimensional node and a two-dimensional boundary becomes a one-dimensional edge. Another feature of IndoorGML is the multi-layered description of indoor spaces. The same space can be divided in cells by rooms, corridors, and staircases; in cells defined by WiFi coverage; or in cells described as restricted and public areas. Each of these layers have their own semantics and NRG.

While IndoorGML focuses on the topological and semantic representation of indoor spaces, it is purposefully limited when modeling interior geometries of these spaces, including objects such as windows, columns, and furniture. This is due to the fact that it was designed to allow integration with other standards such as OGC's previously defined CityGML. As the name suggests, CityGML is meant to be a standard capable of modeling entire cities, at multiple levels of detail (LoD). In CityGML, there exist five LoDs, wherein LoD 0 and LoD 4 are the highest and lowest abstractions, respectively. In particular, LoD 4 allows the geometric representation of building interiors. Ryoo et al. [23] present a comprehensive comparison between the IndoorGML and CityGML LoD 4 standards and their respective strengths and weaknesses. Kim et al. [24] further discuss the integration of the two standards, along with issues and potential solutions.

Here, we study the integration of IndoorGML with SMS data, assuming to be compliant with the MTConnect standard. MTConnect is a semantic standard with controlled vocabulary, types, and relationships for data collected from manufacturing assets. This data could be used to analyze and optimize manufacturing processes or to monitor machines from a workshop floor. The data can contain information like device identity, attributes about its components, and machining parameters, e.g., maximum speeds and axes lengths. Ideally, MTConnect data is captured in near-real-time and collected by *agents*, which share the data via Hypertext Transfer Protocol (HTTP). Stored in a non-proprietary XML-format, MTconnect data streams are read-only, prohibiting command messaging to the monitored machines.

By enriching geospatial data conforming to IndoorGML with near-real-time data conforming to MTConnect, it is possible to create a static representation of a workshop supported by dynamic process data. We created a virtual scene in Unity3D[6] that is capable of leveraging both standards. We expanded the IndoorGML file with additional information, such as the HTTP-address of the MTConnect-stream and position as well as ex-

[6] Available at https://unity.com/

ternal references to the geometrical shape or further documents of objects, e.g., machining centers. The Unity3D-scene creates a 3D representation of the workshop based on the spatial data, placing the defined objects within the scene. Once the scene is rendered, connection to the MTConnect data streams is established and the received data gets processed. This data can now be used to display the current state of each machine. The scene creation is explained in detail within the use cases in Sec. 4.

4 Implementation

To show potential use cases in merging a geospatial description with near-real-time process data, we developed two virtual scenes. These scenes were developed within Unity3D, which is mostly used as a game engine but can also be leveraged for the use in visualization and simulation in other domains. For this prototype, we used the SMS Test Bed[7] as a geospatial representation of a typical workshop. Some of the machines of the SMS Test Bed are connected by the MTConnect standard, which is also used for this prototype. Figure 3 presents a workflow for our two uses cases derived from the same data integration process. As shown in the figure, we anticipate the production of additional use cases through this integration.

4.1 Testing Facility and Setup

The SMS Test Bed represents a digital architecture built on top of a real contract manufacturer at the National Institute of Standards and Technology (NIST). Academic and industrial researchers leverage the SMS Test Bed as both a reference implementation of a set of MTConnect-enabled devices with varying levels of capabilities as well as a rich data source [25]. In our demonstration, we parse near real-time MTConnect data streams from the SMS Test Bed and track the availability of the floor's resources. Note that our prototypes rely on whatever data is available from the NIST shops at the time it is run.

4.2 Use Case 1: Workshop Monitoring

The first use case demonstrates the potential benefits realized by the IndoorGML-MTConnect integration for a supervisor of a workshop. In the case, the supervisor requires a simple way to understand the current status (i.e., machine tool availability) of the floor. With the tendency to implement more automated workflows in manufacturing processes, such a prototype can help to get quick feedback about the production floor's current state.

Once the virtual scene has been initialized, the user selects and loads the modified IndoorGML file. A 3D view of the workshop floor is then rendered displaying all machines defined within the space. The models of the machines are defined by an external reference to an object file (.OBJ) as part of the IndoorGML file. The image section visible to the user is defined by a camera object inside Unity3D. This camera can be moved around freely inside the scene or set to a map-like view.

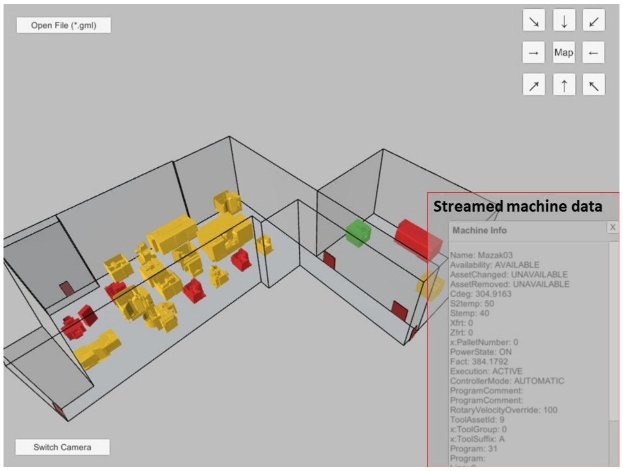
When the connection to the MTConnect stream has been successful, the machines are attributed a color according to their status, e.g., availability status and alarm codes, which in our particular case is set by the status of availability. There are three possible options: available (green), unavailable (red), and unknown (yellow). This helps to quickly get an overview over the whole workshop and identify potential problems easier. Figure 4 shows a 3D view of a workshop with its machines. Since most of the machines in the NIST Shops are not yet digitilized via the SMS Test Bed, many models are colored yellow. Based on the accessible data, it could also be helpful to add the status of current machining process and any occurring errors or unexpected events. When users want to obtain detailed information about one specific machine, they can click on it in the 3D view, which will open a window displaying all data that available through the SMS Test Bed. Another scene feature is the ability to switch the camera into first person view. Users are now able to navigate around the workshop and interact with virtual doors to enter other rooms. While activating this mode, the collision with walls and objects is enabled, so users can move around as workers would be able to in the real world. This is helpful when planning the layout of a workshop and to quickly see if the position and rotation of a machine is useful in terms of productivity, maintainability, and user friendliness.

Despite not being implemented in this prototype, it would be trivial to modify the first person view so that the virtual scene can be leveraged on a head-mounted display in a VR environment. Migrating to AR systems requires more work since synchronization between the virtual and physical spaces is required.

4.3 Use Case 2: Inspection and maintenance

The second use case focuses on aiding maintenance and inspection in production systems. After the modified IndoorGML file and the connected MTConnect data stream have been loaded (akin to the first use case), a map-like view of the workshop is presented to the user. Additionally, the states and transitions are displayed to indicate possible routes around the workshop. As discussed before, IndoorGML supports graph-based representations, wherein each state and transition represents a node and edge, respectively. This is especially useful for informing indoor navigation. To demonstrate this point, our prototype allows users to select a way-point or a machine, either by clicking on it or choosing it from a drop-down list, as a start and endpoint. Since the system of way-points, paths and their length is known, we leverage Dijkstra's algorithm to calculate the shortest path between the start and endpoint. The calculated path and its total length is displayed (see Fig. 5). This feature enables mainte-

[7] Available at https://smstestbed.nist.gov

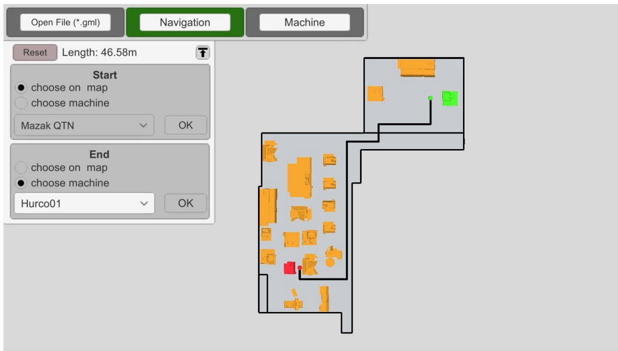


Streamed machine data

Machine Info

Name: Mazak03
Availability: AVAILABLE
AssetChanged: UNAVAILABLE
AssetRemoved: UNAVAILABLE
Cdeg: 304.9163
S2temp: 50
Stemp: 40
Xfrt: 0
Zfrt: 0
x:PalletNumber: 0
PowerState: ON
Fact: 384.1792
Execution: ACTIVE
ControllerMode: AUTOMATIC
ProgramComment:
ProgramComment:
RotaryVelocityOverride: 100
ToolAssetId: 9
x:ToolGroup: 0
x:ToolSuffix: A
Program: 31
Program:

Figure 4. 3D view of a workshop with machines colored based on availability. MTConnect stream of chosen machine is displayed on the right.



Length: 46.58m

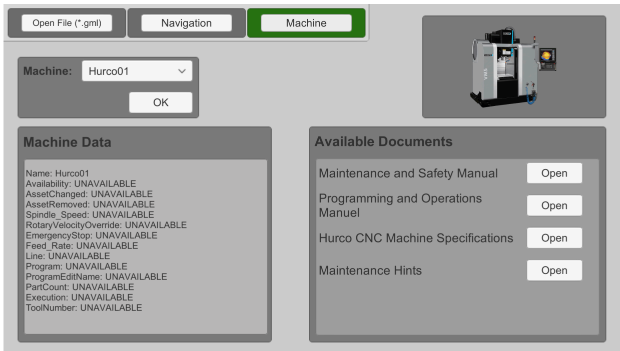
Start
● choose on map
○ choose machine
Mazak QTN
OK

End
○ choose on map
● choose machine
Hurco01
OK

Figure 5. View of the calculated path between two way-points

nance workforce to quickly find the machines to which must be attended, if they are not familiar with the layout of the workshop.

For the purpose of maintenance and inspection, it is helpful to procure further details about the machine once the worker has reached it. Therefore, we created an additional tab labeled ”Machine” (see Fig. 6). This enables users to choose a machine from a drop-down list to obtain an overview outlining the available information according to their choice. Our prototype shows live data available via MTConnect, pictures of the machine, and additional documents akin to a handbook, programming manual, or hints for maintenance.



Machine: Hurco01

Machine Data

Name: Hurco01
Availability: UNAVAILABLE
AssetChanged: UNAVAILABLE
AssetRemoved: UNAVAILABLE
Spindle_Speed: UNAVAILABLE
RotaryVelocityOverride: UNAVAILABLE
EmergencyStop: UNAVAILABLE
Feed_Rate: UNAVAILABLE
Line: UNAVAILABLE
Program: UNAVAILABLE
ProgramEditName: UNAVAILABLE
PartCount: UNAVAILABLE
Execution: UNAVAILABLE
ToolNumber: UNAVAILABLE

Available Documents

Maintenance and Safety Manual
Programming and Operations Manuel
Hurco CNC Machine Specifications
Maintenance Hints

Figure 6. Detailed view of machine information via the “Machine” tab

4.4 Limitations of Use Cases

Within the current state of our software prototype, there are limitations which can be tackled through future work. With respect to the representation of the shop floor, we currently only place machine tools into the virtual scene. In reality, there are additional objects that could be considered, including static objects, such as desks and cabinets, as well as dynamic objects, such as humans, carts and AGVs.

Considering whether an object is affixed to some coordinates or has the ability to move is handled through CityGML. Note that we have not yet fully implemented this integration aspect but plan to in the near future. As Fig 7 suggests, `MTConnectAssets` can be modeled as `BuildingFurniture` within CityGML. Movable entities are meant to be modeled using the `BuildingFurniture` element. The static objects, i.e., `MTConnectDevices`, could follow the modeling paradigm of the `BuildingInstallation` element defined within CityGML. With objects that are dynamical in position, finding the indoor location of these objects and transferring those positions to the Unity3D scene are both required. Such extensions would especially be helpful for tracking commonly misplaced objects on the floor, such as tooling and measurement devices.

Another limitation lies within the digital models themselves. First of all, the solid models of the machines are separated from the modified IndoorGML file. Even though this separation allows for modification of the models without affecting the geospatial representation, this solution is not a fully integrated one. In the future, we plan to consider embedding these models within the CityGML description of objects within the indoor space. The `lod4Geometry` element of CityGML offers native support for the description of object geometry. Additionally, the used 3D models are rigid and their geometry has been simplified, so that their shape can be described with fewer polygons, which led to a better performance of our prototype. Their purpose in our prototype was to leverage these models to recreate a quick, decently realistic layout of the existing workshop. If there exist emerging needs to display the movement of digital assemblies, the models must be split into multiple objects based on their independently

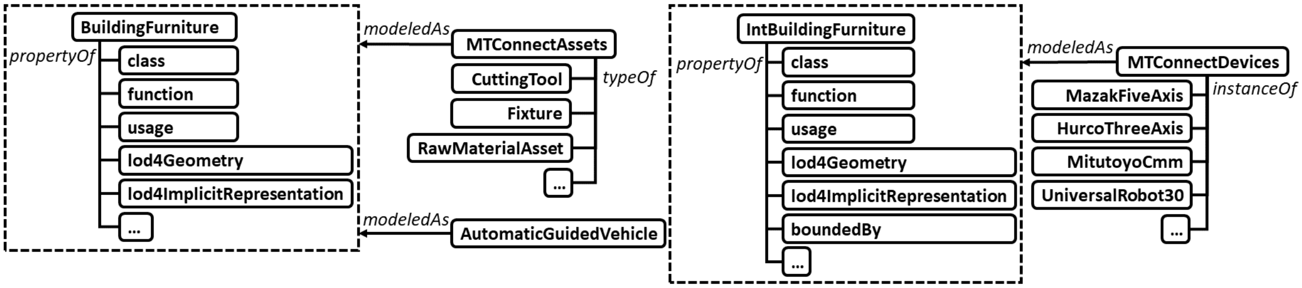


Figure 7. Elements and attributes represented in the CityGML object types critical to our integration. (A) `BuildingFurniture` can be used to model production floor tooling and other equipment that is expected to change its positional data often. (B) `IntBuildingInstallation`, traditionally used to model columns and other practically immovable objects, is used to model large production devices, e.g., CNC machining centers.

moving sub-assemblies and then realigned based on their position and kinematic properties.

When the layout of the workshop is extended over multiple floors, it is still possible to display that within our prototype since IndoorGML offers multi-floor modeling. However, in our prototype, extending to a multi-floor instance could cause user interaction and experience issues, e.g., occlusion challenges. To include machine-specific data within our modified IndoorGML file, significant manual labor is required. These activities could be (semi-)automated by either (1) formally appending the CityGML and IndoorGML with data elements designed to capture production-specific scenarios or (2) mapping similar elements that already exist in these data structures with knowledge graphs. Improving the automation of such procedures would expedite the initial process of implementing our prototype in other workshop environments. With more formal methods deployed, it would also be possible to develop an editor tool to visually construct our modified spatial representation on the fly. Additionally, a formal amendment to the schema representation would allow for proper governance and validation procedures.

5 Extensions to other technologies

Based on the research opportunities described in the previous section, we describe the technology implications for two domains: augmented reality and automated guided vehicles. Moving forward, we plan to test these geospatial representations and concepts to meet challenges faced by both domains.

5.1 Implications for Augmented Reality

Current AR systems suffer from interoperability issues caused by a lack of a shared coordinate system in which to operate. This is made obvious when trying to use different devices, e.g., tablets and head-mounted displays, AR frameworks, e.g., Vuforia, ARKit, and ARCore), or tracking techniques, e.g., marker-based tracking or marker-less tracking. Such techniques and technologies are often incompatible with each other in many ways, including the understanding of the space in which they operate. For instance, a marker-less AR application built using Apple's iOS exclusive framework ARKit[8] might place a virtual object in space relative to a feature map that would be incompatible with Google's ARCore[9] toolkit. Moreover, a marker-based AR application would not be able to make sense of the position of the virtual object since it would not track the features to which it is relatively positioned. To this end, most industrial AR installations exist as isolated use-case-specific prototypes that are only *aware* of their immediate surroundings rather then the larger context of the manufacturing floor. As a result, most AR applications are unable to share spatial information consistently.

While infeasible for many AR applications, in environments that can be easily predefined, such as manufacturing floors, standard geospatial representations can help define a shared underlying coordinate system on top of which applications can be built. This would lead to a better integration of industrial AR and virtual reality (VR) systems regardless of the devices or frameworks used. This notion has garnered attention from the standards development community. Recently, the OGC has chartered a new standards effort for representing device and camera poses with respect to global coordinates. We hope these efforts will realize a new standard representation that is already consistant with existing OGC standards, i.e., CityGML and IndoorGML.

5.2 Implications for Automated Guided Vehicles

Automatic (or automated[10]) guided vehicles (AGVs) presents another domain for which geospatial representations can

[8] Available at https://developer.apple.com/augmented-reality/arkit

[9] Available at https://developers.google.com/ar

[10] In this paper, we consider automatic guided vehicles and autonomous vehicles, e.g., mobile robots, as the same. To learn more about their distinction, refer to ASTM Committee F45. F45 addressed issues related to performance standards and guidance materials for AGVs. More information is available at https://www.astm.org/COMMITTEE/F45.htm.

help performance. Clearly, AGVs already have on-board capabilities for constructing ad-hoc maps of indoor spaces and avoiding obstacles. However, there still remain challenges related to the lack of semantically rich data available to AGVs. For example, consider a robotic arm mounted on an AGV tasked to approach a table and engage with some physical object. Without formally encoding that the AGV is allowed to get very near, the AGV could stop short of the target table in fear of collision. This simple example demonstrates the importance of formally preserving the context of the physical environments. The target area should be treated differently than walls, columns, or other hazards. We envision that the integration of point cloud-maps built by AGVs with semantically rich geospatial representations would enhance their mobility and navigation.

Additionally, relating other information standards with such geospatial representations, as we have demonstrated here with the MTConnect standard, can further increase the information available to AGVs at any given time. Looking forward, achieving this kind of integration would enhance the situational awareness of AGVs for better planning and routing purposes. For example, an AGV could move and perform tasks automatically based on machine availability and readiness captured by near-real-time MTConnect data streams. Theoretically, this could enhance the agility and flexibility of a fleet of AGVs servicing tasks related to MTConnect-enabled devices, including loading, unloading, and material transport.

6 Closing & Looking forward

The full realization of smart manufacturing systems remains a work-in-progress. We believe that standardizing spatial data representations for production systems is vital for fully describing factories digitally. To this end, we presented a prototype that makes use of existing standards to combine geospatial descriptions (via IndoorGML) and real-time process data (via MTConnect) of machines. We set up two different use cases, focusing on both monitoring and navigation, to demonstrate the usefulness of merging these two areas.

Further development should consider creating a framework that supports the the combination of multiple sources of information that already exist in modern factories. This can include positions of workers and tools, RFID-sensor readings, video feeds of surveillance cameras, or energy consumption measurements. By bundling such information, the development of applications which support the control of smart factories can be developed much easier since the acquisition of data would be standardized. In other words, from a design perspective, leveraging standards for representing geospatial and device data in production systems facilitates more flexible prototype interface development. The more consistently information is represented, the simpler development of new applications becomes.

Furthermore, we see significant research opportunities in sensor fusion for more precise geospatial alignment. One example is leveraging on-board sensors from AGVs and more contextually defined, static geospatial definitions, such as those offered by IndoorGML and CityGML. If successfully integrated, such definitions could enable safer AR installations. For example, adding context to a 2-D point cloud delivered by an AGV could help avoid dangerous occlusion problems, e.g., rendering a digital object and effectively blocking view of a safety hazard. We are currently planning a partnership with the Robotics Program at NIST to investigate these research topics.

DISCLAIMER

This work represents an official contribution of NIST and hence is not subject to copyright in the US. Identification of commercial systems are for demonstration purposes only and does not imply recommendation or endorsement by NIST.

ACKNOWLEDGEMENTS

We thank Christine Perey, co-founder of the Augmented Reality for Enterprise Alliance (AREA), a consortium focused on industrial AR, for her valuable input. We also acknowledge the NIST-internal as well as the anonymous reviewers for their helpful comments.

REFERENCES

[1] Helu, M., and Hedberg Jr, T., 2015. “Enabling smart manufacturing research and development using a product lifecycle test bed”. *Procedia manufacturing,* ***1***, pp. 86–97.

[2] Burke, R., Mussomeli, A., Laaper, S., Hartigan, M., and Sniderman, B., 2017. “The smart factory: Responsive, adaptive, connected manufacturing”. *Deloitte Insights,* ***31***(1), pp. 1–10.

[3] Thoben, K.-D., Wiesner, S., and Wuest, T., 2017. ““industrie 4.0” and smart manufacturing-a review of research issues and application examples”. *International journal of automation technology,* ***11***(1), pp. 4–16.

[4] Lu, Y., Morris, K. C., and Frechette, S., 2016. “Current standards landscape for smart manufacturing systems”. *National Institute of Standards and Technology, NISTIR,* ***8107***, p. 39.

[5] Beinschob, P., and Reinke, C., 2015. “Graph slam based mapping for agv localization in large-scale warehouses”. In 2015 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, pp. 245–248.

[6] Lee, S., and Akin, Ö., 2011. “Augmented reality-based computational fieldwork support for equipment operations and maintenance”. *Automation in Construction,* ***20***(4), pp. 338–352.

[7] Hao, Y., and Helo, P., 2017. "The role of wearable devices in meeting the needs of cloud manufacturing: A case study". *Robotics and Computer-Integrated Manufacturing,* ***45***, pp. 168–179.

[8] Scholz, J., and Schabus, S., 2017. "Towards an affordance-based ad-hoc suitability network for indoor manufacturing transportation processes". *ISPRS International Journal of Geo-Information,* ***6***(9), p. 280.

[9] Nee, A. Y., Ong, S., Chryssolouris, G., and Mourtzis, D., 2012. "Augmented reality applications in design and manufacturing". *CIRP annals,* ***61***(2), pp. 657–679.

[10] Padovano, A., Longo, F., Nicoletti, L., and Mirabelli, G., 2018. "A digital twin based service oriented application for a 4.0 knowledge navigation in the smart factory". *IFAC-PapersOnLine,* ***51***(11), pp. 631–636.

[11] Barberá, H. M., Quinonero, J. P. C., Izquierdo, M. A. Z., and Skarmeta, A. G., 2003. "I-fork: a flexible agv system using topological and grid maps". In 2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422), Vol. 2, IEEE, pp. 2147–2152.

[12] Johnston, A. R., Assefi, T., and Lai, J. Y., 1979. "Automated vehicle guidance using discrete reference markers". *IEEE Transactions on Vehicular Technology,* ***28***(1), pp. 95–106.

[13] Ray, A. K., Gupta, M., Behera, L., and Jamshidi, M., 2008. "Sonar based autonomous automatic guided vehicle (agv) navigation". In 2008 IEEE International Conference on System of Systems Engineering, IEEE, pp. 1–6.

[14] Isikdag, U., Zlatanova, S., and Underwood, J., 2013. "A bim-oriented model for supporting indoor navigation requirements". *Computers, Environment and Urban Systems,* ***41***, pp. 112–123.

[15] Kang, H.-K., and Li, K.-J., 2017. "A standard indoor spatial data model—OGC IndoorGML and implementation approaches". *ISPRS International Journal of Geo-Information,* ***6***(4), p. 116.

[16] Goodrum, P. M., McLaren, M. A., and Durfee, A., 2006. "The application of active radio frequency identification technology for tool tracking on construction job sites". *Automation in construction,* ***15***(3), pp. 292–302.

[17] Schabus, S., and Scholz, J., 2015. "Geographic information science and technology as key approach to unveil the potential of industry 4.0: How location and time can support smart manufacturing". In 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), Vol. 2, IEEE, pp. 463–470.

[18] Carrasco, U., Coronado, P. D. U., Parto, M., and Kurfess, T., 2018. "Indoor location service in support of a smart manufacturing facility". *Computers in Industry,* ***103***, pp. 132–140.

[19] Ganesharajah, T., Hall, N. G., and Sriskandarajah, C., 1998. "Design and operational issues in agv-served manufacturing systems". *Annals of operations Research,* ***76***, pp. 109–154.

[20] Teichholz, E., 2004. "Bridging the aec/fm technology gap', ifma facility management journal". *March/April*.

[21] Lee, S., and Akin, Ö., 2009. "Shadowing tradespeople: Inefficiency in maintenance fieldwork". *Automation in Construction,* ***18***(5), pp. 536–546.

[22] Palmarini, R., Erkoyuncu, J. A., Roy, R., and Torabmostaedi, H., 2018. "A systematic review of augmented reality applications in maintenance". *Robotics and Computer-Integrated Manufacturing,* ***49***, pp. 215–228.

[23] Ryoo, H.-G., Kim, T., and Li, K.-J., 2015. "Comparison between two ogc standards for indoor space: Citygml and indoorgml". In Proceedings of the Seventh ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, pp. 1–8.

[24] Kim, J.-S., Yoo, S.-J., and Li, K.-J., 2014. "Integrating indoorgml and citygml for indoor space". In International Symposium on Web and Wireless Geographical Information Systems, Springer, pp. 184–196.

[25] Helu, M., Hedberg Jr, T., and Feeney, A. B., 2017. "Reference architecture to integrate heterogeneous manufacturing systems for the digital thread". *CIRP journal of manufacturing science and technology,* ***19***, pp. 191–195.

Correcting Systematic Energy Deficits in the Laser-Pulsed Atom Probe Mass Spectrum of SiO_2

Benjamin W. Caplins[1], Paul T. Blanchard[2], Ann N. Chiaramonti[1], David R. Diercks[2,3], Luis Miaja-Avila[2], Norman A. Sanford[2]

[1]National Institute of Standards and Technology, Applied Chemicals and Materials Division, Boulder, CO, 80305, United States
[2]National Institute of Standards and Technology, Applied Physics Division, Boulder, CO, 80305, United States
[3]Colorado School of Mines, Metallurgical and Materials Engineering, Golden, CO, 80401, United States

Voltage-pulsed atom probe tomography is difficult or impossible to use for materials with poor electrical conductivity. Laser-pulsed atom probe tomography has been shown to overcome this limitation by triggering the field evaporation optically and enables the analysis of strongly insulating materials[1]. However, using a laser-pulse to trigger ion evaporation does not completely render the electrical insulating properties of the tip irrelevant. For example, when the electrical resistivity of the tip is high enough, the atto-amp to femto-amp ion current emitted from the tip apex can result in a significant voltage drop along the tip axis (Figure 1a)[2]. This voltage drop has the effect of giving the emitted ions an `energy deficit' – i.e. the actual accelerating voltage for the ion's is smaller than the voltage applied experimentally to the base of the sample. A static energy deficit will be resolved when a mass spectrum is calibrated, however, when the voltage drop fluctuates throughout an atom probe dataset it cannot be corrected with any standard analysis methods. This effect has been noted in the literature for infrared, visible, and ultraviolet laser wavelengths[3]. More recently it also showed up when using extreme ultraviolet (EUV) light[4] to triggered field ion evaporation of SiO_2 -- Figure 1b shows a time-of-flight history for an SiO_2 tip held at constant voltage collected using an pulsed 29.6 nm light source. Over the course of the run, the pointing and intensity of the EUV light source fluctuated resulting in a changing ion current and therefore a changing energy deficit which significantly degrades the quality of the resulting mass spectrum (Figure 1f) which would degrade subsequent analyses.

In order to correct the data for this unknown (fluctuating) acceleration voltage we have developed an empirical algorithm. The algorithm first splits the data into a large number of contiguous `chunks'. Then, the chunks are optimally aligned to a reference spectrum with a multiplicative correction factor (Figure 1c) using a histogram-based dot product optimization metric (Figure 1e). Applying a log-transform to the time-of-flight data enables the use of the FFT for the optimization step. The result is that a correction factor can be determined for every ion's flight time that minimizes the effect of the unknown (varying) voltage drop (Figure 1d). The resulting mass spectrum of the corrected data is significantly improved (Figure 1f) versus the uncorrected data. This correction algorithm ensures that the highest quality input data is supplied into the atom probe data analysis pipeline. In principle, this method can be extended to spatially heterogenous materials using a multireference approach which would be facilitated by well-known unsupervised machine learning algorithms.

Caplins, Benjamin; Blanchard, Paul; Chiaramonti Debay, Ann; Diercks, David; Miaja Avila, Luis; Sanford, Norman A. "Correcting Systematic Energy Deficits in the Laser-Pulsed Atom Probe Mass Spectrum of SiO2." Paper presented at Microscopy and Microanalysis 2020, Milwaukee, WI, US. August 02, 2020 - August 06, 2020.

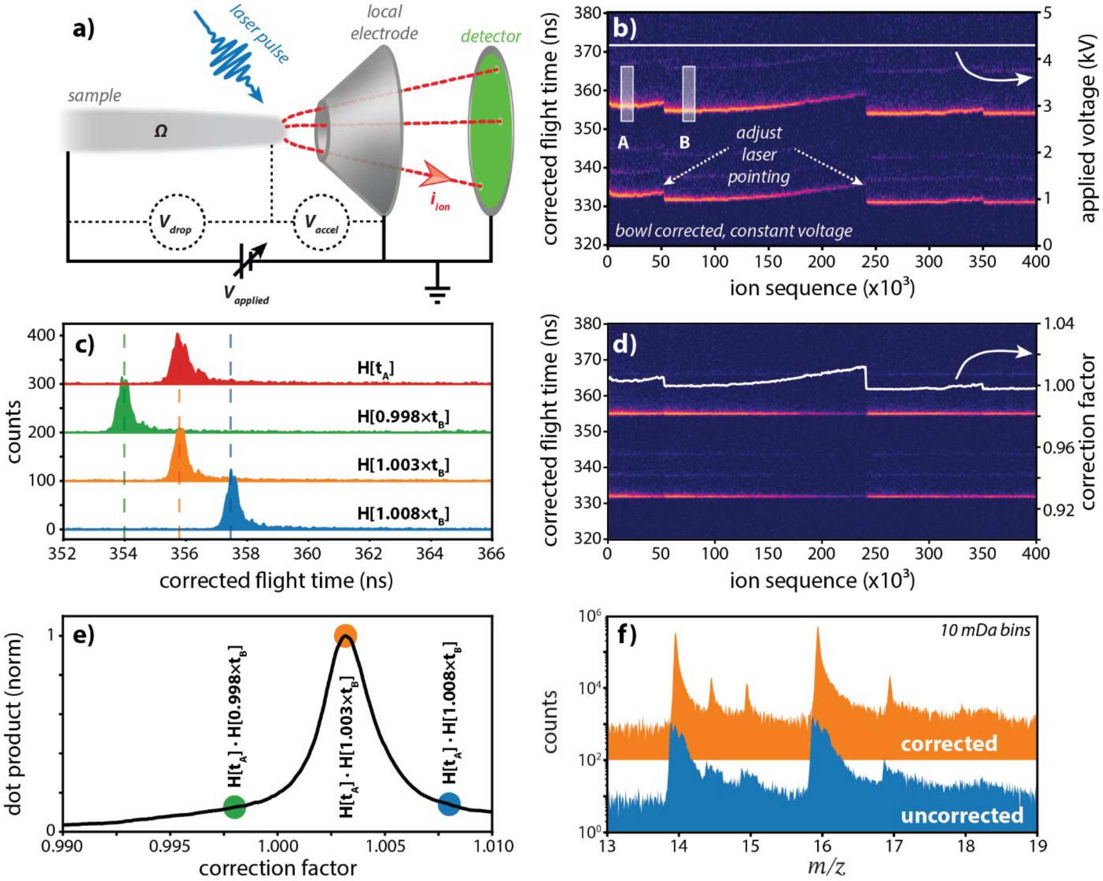


Figure 1: (a) Schematic of a laser-pulsed atom probe experimental setup. (b) Time-of-flight history of a constant voltage SiO_2 experiment. (c-e) An empirical algorithm corrects for the unknown (time-varying) voltage drop. (d) The mass spectrum generated from the corrected data has significantly improved quality.

References:

[1] B. Gault et al., *Appl. Phys. Lett.* **88** (2006), p. 114101.
[2] L. Arnoldi et al., *J. Appl. Phys.* **115** (2014), p. 203705.
[3] L. Arnoldi et al., *J. Appl. Phys.* **126** (2019), p 045710.
[4] A. Chiaramonti et al., *MRS Adv.* **4** (2019), p. 2367.
[5] This work is a contribution of the US Government and is not subject to United States copyright.

HYBRID MODELING APPROACH FOR MELT POOL PREDICTION IN LASER POWDER BED FUSION ADDITIVE MANUFACTURING

Tesfaye Moges[1], Zhuo Yang[2], Kevontrez Jones[3], Shaw Feng[4], Paul Witherell[4], Yan Lu[4]

[1] Guest Researcher at National Institute of Standards and Technology (NIST) and Affiliated with Indian Institute of Technology Delhi, New Delhi, India
[2] Guest Researcher at NIST and Affiliated with University of Massachusetts, Amherst, MA, USA
[3] Northwestern University, Evanston, IL, USA
[4] National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

ABSTRACT

Multi-scale multi-physics computational models are a promising tool to provide detailed insights to understand the process-structure-property-performance relationships in additive manufacturing (AM) processes. To take advantage of the strengths of both physics-based and data-driven models, we propose a novel hybrid modeling framework for laser powder bed fusion (L-PBF) processes. Our unbiased model integration method combines physics-based data and measurement data for approaching more accurate prediction of melt pool width. Both a high-fidelity computational fluid dynamics (CFD) model and experiments utilizing optical images are used to generate a combined dataset of melt pool widths. From this aggregated dataset, a hybrid model is developed using data-driven modeling techniques, including polynomial regression and Kriging methods. The performance of the hybrid model is evaluated by computing the average relative error and compared with the results of the simulations and surrogate models constructed from the original CFD model and experimental measurements. It is found that the proposed hybrid model performs better in terms of prediction accuracy and computational time. Future work includes a conceptual introduction on the use of an AM ontology to support improved model and data selection when constructing hybrid models. This study can be viewed as a significant step towards the use of hybrid models as predictive models with improves accuracy without the sacrifice of speed.

Keywords: Additive manufacturing, Laser powder bed fusion, Hybrid model, Melt pool width, Gaussian process/Kriging, Data-driven surrogate model, Ontology

1. INTRODUCTION

Metal additive manufacturing (AM) produces metallic parts by fusing materials in a layer-by-layer fashion directly from a 3D CAD model [1]. Laser powder bed fusion (L-PBF) is the most common AM process used for the fabrication of metallic components. In the L-PBF process, a thin powder layer is spread on a substrate and a laser beam selectively melts and fuses powder with neighboring particles and previous layer. This process is repeated until the final part is formed. L-PBF has tremendous potential of producing metallic parts with complex geometry, internal structures, and conformal heating/cooling channels for a wide range of applications including aerospace, automotive, and biomedical implants [2–4]. Compared to conventional manufacturing techniques, L-PBF has many advantages as it allows to locally control microstructures by varying process parameters so that parts with desirable mechanical properties can be produced [5,6]. It also minimizes material wastes and reduces lead time [7]. While offering many advantages, the L-PBF process also faces challenges such as inconsistent part quality and defects in terms of porosity, poor surface finish, delamination, crack formation, and residual stress [8]. Without proper control mechanisms, these challenges can lead to unstable mechanical properties and poor dimensional accuracy on produced parts.

To overcome these challenges and help develop process control mechanisms, there have been many research efforts that have aimed to understand the influence of different process parameters and material properties on part quality [9,10]. The research efforts can be broadly categorized as experimental-based and physics-based investigations. The experimental-based investigation is more realistic as it directly captures observed physical phenomena occurring during the process. However, this

Moges, Tesfaye; Yang, Zhuo; Jones, Kevontrez; Feng, Shaw C.; Witherell, Paul; Lu, Yan. "Hybrid Modeling Approach for Melt Pool Prediction in Laser Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

approach can be time consuming and costly while lending itself to inherent process variabilities and sensor calibration errors that cause large measurement uncertainty [11]. Additionally, not every process parameter can be measured. For these reasons, amongst others, interest in computational models for AM continues to rise.

For the L-PBF process, computational models have been widely used to simulate the heat transfer, fluid flow, and phase transformations in L-PBF process and estimate the temperature fields, flow velocities, melt pool characteristics, solidification rate, and residual stress [12,13]. Although these computational models are promising tools to understand the physics of the process, their prediction accuracy can be potentially affected by the assumptions used during model formulation [14]. In addition to modeling assumptions, inaccurate selection of model parameters, such as absorption coefficient and material properties, also leads to significant discrepancy between computationally predicted and experimental results [15]. Furthermore, even though high fidelity physics-based models describe the L-PBF process in detail, solving these models is time consuming due to the complex physics.

Overall, physics-based models suffer from prediction accuracy caused by uncertainties that are associated with the modeling assumptions made about the L-PBF process. The prediction accuracy of the experimental-based investigations is highly affected by sensor noise and measurement uncertainties due to the inherent complexity and stochastic nature of the process. Alterations to each approach have sought to compensate for their drawbacks. For instance, reduced-order physics-based models have been sought to reduce computational demands, though the tradeoff is they can be less accurate. Highly complex, multi-variable experiments of processes like L-PBF are also difficult to implement due to their high cost and time requirements. We propose that the most effective way to overcome the challenges associated with each approach is to develop a model that embraces the advantages of each approach.

In this study, we propose a hybrid model by intertwining physics-based simulation data and experiment-based empirical data for the prediction of output quantities of interest (QoIs) in L-PBF process. To develop an unbiased model which integrates computational and experimental data, data-driven modeling techniques including polynomial regression and Gaussian process/Kriging methods, which are suitable for multi-dimensional problems having limited data, are employed. Melt pool width is chosen as our primary output QoI since the melt pool plays a significant role in determining the microstructure, residual stresses, and mechanical properties of a part fabricated by L-PBF process [16]. A CFD model is used to predict melt pool widths for various combinations of laser power and scan speed. Empirical data is obtained from ex-situ melt pool width measurement data taken from single-track experiments with similar process parameters.

In developing the hybrid model, a polynomial regression method is first applied to construct the initial simulation-based surrogate model using data from a CFD model. Then, the Kriging method is applied to model the residual error between experimental and computational results. An adaptive modeling method is used to iteratively update the Kriging model to improve the predictive errors of the surrogate model. to improve the model, measurement data are iteratively selected by updating sample points using the maximum average relative error as a determining factor. The performance of the hybrid model is evaluated by comparing the relative error to the individual simulation-based and experiment-based surrogate models. It is found that the hybrid model performs better in terms of prediction accuracy and computational time. Once the hybrid model is developed, making predictions on a new set of input variables is straightforward.

This paper is organized as follows: in Section 2, we briefly review the prior research efforts on the physics-based computational models focusing on thermal models and data-driven modeling approaches used in L-PBF AM. Then we provide an overview of the CFD model, the data-driven modeling techniques used in this study, and the hybrid model development methodology in Section 3. In Section 4, we demonstrate the proposed methodology to develop a hybrid model for melt pool width prediction and analyze its performance using CFD simulation data and measurement data. A conceptual use of AM ontology for model selection is introduced in Section 5. Section 6 presents the concluding remarks of the current study and future work. We view this work to be an essential step to developing fast and accurate hybrid models that compliment both physics-based and experimental approaches to improve L-PBF part quality.

2. BACKGROUND

Though L-PBF has shown to be a capable technology for producing complex parts, challenges remain, e.g., inconsistent part properties with defects. Part defects are often attributes to porosity in the fabricated parts, surface roughness, anisotropy in microstructure, residual stresses, delamination, and cracks [17]. Physics-based modeling techniques are based on cause-effect principles in physics, such as fluid dynamics, thermodynamics, heat transfer, and kinetics. While such techniques provide invaluable insight into the general behavior of a process, fall short in predicting defect generation and propagation and final part quality for a specific process implementation. There are at least three reasons for that: (a) a complete set of process variables may not be obtainable due to a lack of knowledge regarding the process, (b) unknown complex energy and material behaviors in the process, such as non-linear interactions, and (c) the physics-based models have not been validated due to limitation in available measurement techniques.

AM data analytics includes measured data, data fusion, data analysis, statistical methods, and machine learning [18]. Real data are acquired from in-situ or ex-situ sensors. To model the process, the relations between inputs and outputs are created by data-driven modeling techniques. Data-driven models can be employed to model the non-linear relationship between sensor outputs with process parameters and predict output quantities of interest for different sets of input variables. However, as stated previously, the measurement data used to build the data-driven models are affected by uncertainties related to error in sensor calibration and noise, imprecise measurement methods, and variations in the measurements. The following subsections review relevant works associated with each of these modeling approaches.

2.1. Physics-based thermal models

There have been tremendous efforts in the last decade in developing computational models to simulate powder layer deposition, powder-laser interactions, melt pool formation, solidification and grain growth, and residual stress and deformation with different levels of fidelity in the L-PBF process [12–14,19]. The term fidelity used in this context is based on the different physical phenomena captured by the computational models. For instance, a model that captures a larger number of phenomena refers as high-fidelity, whereas, low fidelity is the one with least number of phenomena.

Physics-based computational modeling has been crucial to understanding process-structure-property relations in metal AM [20], and these models have come in different varieties. With regard to the transient manufacturing process, thermal models based on the semi-analytical, the finite element, and finite volume methods have been developed. The semi-analytical Rosenthal-based low-fidelity thermal models can solve the heat conduction equation for temperature profile and melt pool geometry [21,22]. However, they neglect physical phenomena such as the effective powder layer, laser spot diameter, and other phenomena related to heat transfer and fluid flow. Typically, the finite element medium-fidelity models can provide a thermal history of an entire part being built [23–26], but they do so by purely considering heat conduction and neglecting the fluid flow behavior within the melt pool. This simplification can lead to predictions of inaccurate temperature fields. For example, Manvatkar et al [27] showed that by ignoring Marangoni convection in the molten melt pool, cooling rates in laser assisted AM may be overestimated by as much as double the correct values. Conversely, Gan et al [28] demonstrated that incorporation of fluid-flow and vaporization can significantly enhance a model ability to accurately predict melt pool geometry, peak temperature and surface topology. For these reasons, several high-fidelity models based on the finite volume method and CFD have been developed to account for additional physics within the melt pool [29–34].

Ensuring the accuracy of high-fidelity models requires an extensive use of well-designed and highly controlled experiments for validation. Due to extremely high temperatures, violent metallic powder spattering, and highly complex physics occurring at multiple length scales within very short time scales, it is difficult to conduct in-situ measurements in the L-PBF process. Typically, as is the case for Ghosh et al [35], ex-situ measurements are used for numerical validation. It is apparent that assumptions and simplifications made about the process can significantly affect a computational model's predictive capabilities.

2.2. Data-driven models

Data-driven modeling methods, as a more stochastic approach, have been deployed to analyze AM data. Different data-driven approaches have been implemented to construct surrogate models from experiment-based empirical data and physics-based simulation data. Data-driven surrogate models that rely on empirical data can help estimate data points in a new design space and evaluate correlations between input parameters and output QoIs [36,37]. In addition, simulation-based data-driven surrogate models aim support a black-box approach to reduce the high computational cost of high-fidelity simulations [38]. Data-driven machine learning techniques have been used in AM for different applications throughout the AM lifecycle [39].

Using experiment-based empirical data, Fathi and Mozaffari [40] developed a data-driven framework to relate process parameters including laser power, scan speed, and powder flow rate to melt pool depth and deposition height for the laser-based direct energy deposition (DED) process. Similarly, Lu et al. [41] used a neural network to map process parameters to deposition height for the DED process. Since obtaining measurement data for such complex processes is difficult and time consuming, in lieu of empirical data, researchers have used physics-based simulation data to develop data-driven surrogate models. From the semi-analytical Rosenthal-based thermal model, Yang et al. [42] developed a Dynamic Variance-Covariance Matrix (DVCM) method to investigate the influence of input parameters such as laser power, scan speed, absorption coefficient, and thermal diffusivity on melt pool width for L-PBF process. Kamath [43] built data-driven surrogate models from the Eagar-Tsai thermal model using regression trees and Gaussian process regression to predict melt pool depth for the L-PBF process. Recently, Tran and Lo [44] developed an approach to optimize process parameters such as laser power, scan speed, and layer thickness using artificial neural network (ANN) for L-PBF process.

The previous works focused on developing data-driven surrogate models based on either experiment-based empirical data or physics-based simulation data to predict output QoIs. As measurement data contain errors and uncertainties associated with sensor calibration and noise, imprecise measurement methods, and variation in the process; and physics-based simulation data have uncertainty associated with modeling assumptions, numerical approximation, and variability in input parameters, data-driven surrogate models that are developed solely from empirical data or simulation data may exhibit variable predictive capabilities. Therefore, hybrid models that embraces the advantages of experiment-based and simulation-based approaches are needed.

In general, physics-based models typically do an acceptable job at predicting output QoIs in a wide range of input variables. However, there are multiple sources of uncertainty in these computational models that can cause significant prediction errors [15]. On the other hand, data-driven modeling techniques often rely on a specific system to help find the relationships between inputs and outputs, and without the explicit knowledge of the physical behavior of the system [45]. In such scenarios,

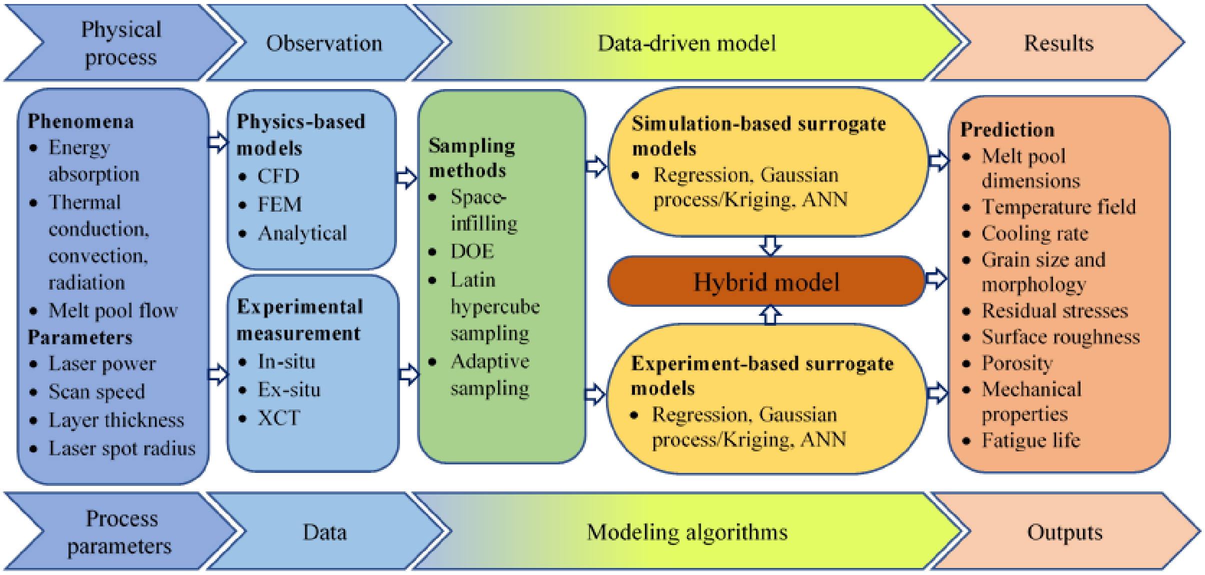


Figure 1. Hybrid modeling framework

due to data sparsity, extrapolation is limited and careful selection of training data, training algorithm, and model complexity is required [46]. The proposed hybrid model offers an approach where extrapolation can be done for a wide input of parameter ranges while capturing the complex and highly non-linear behavior of the L-PBF process [46].

3. HYBRID MODEL DEVELOPMENT

As mentioned above, researchers have developed various approaches by utilizing different scales, disciplines, and perspectives to overcome the complexity of AM processes. All these efforts can significantly improve knowledge mining in AM. However, these efforts also raise more challenges in data filtering, algorithm selection, and model integration. Different simulations, for example, may only work under specific conditions if they were developed based on different physical phenomena. It is also hard to guarantee consistency between different datasets from various sources. Another typical issue is the stochastic error observed between computational and experimental data. This error may initiate from the fundamental hypothesis or numerical approach of the simulation. However, physical experiments cannot benefit from selective simplifying assumptions like simulations. Hence, experimental results tend to include more complicated physical phenomena than those obtained from a simulation. Thus, even high-fidelity simulations cannot truly match the real experimental data based on different AM machines, labs, and material.

Modifications applied to physics-based models can potentially address the above issues. This section presents the development of an unbiased model integration method to combine computational and experimental data to provide accurate predictions regardless of the sample size and fidelity of the data. The high-fidelity physics-based model, though requires higher computational cost, has the freedom to generate a vast amount of data. On the other hand, the experimental data is assumed to be ground truth but usually limited and expensive to sample.

Figure 1 shows the proposed hybrid modeling framework for the L-PBF process. Various combinations of process parameters, physics-based models and experimental measurement techniques provide datasets with inherent uncertainties [47–49]. Then, multiple sampling methods filter through each dataset to systematically create subsets of the data to be used for training and validation. Once these data subsets are determined, they are used in conjunction with data-driven models to build simulation-based and experiment-based surrogate models. To improve the accuracy of these surrogate models, a hybrid model that combines the physics-based data and experimentally measured data using an unbiased model integration method is created. With this framework in mind, a brief overview of the physics-based CFD model and data-driven approaches used in the current study is provided, and the workflow and algorithm of the proposed hybrid model are discussed in detail in this section.

3.1. Brief details on CFD model

A well tested, three dimensional, transient, thermal-fluid flow model for L-PBF [28,29] is adapted to compute temperature and velocity fields to generate the data referenced in this work. The thermal-fluid flow model solves for conservation of mass (Eqn. 1), momentum (Eqn. 2), and energy (Eqn. 3) to consider liquid flow within the melt pool driven by Marangoni convection.

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u_i}{\partial x_i} = 0 \tag{1}$$

Moges, Tesfaye; Yang, Zhuo; Jones, Kevontrez; Feng, Shaw C.; Witherell, Paul; Lu, Yan. "Hybrid Modeling Approach for Melt Pool Prediction in Laser Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

$$\frac{\partial \rho u_i}{\partial t}+\frac{\partial \rho u_i u_j}{\partial x_j}=\frac{\partial}{\partial x_j}\left(\mu\left(\frac{\partial u_i}{\partial x_j}\right)\right)-\frac{\partial p}{\partial x_i} -\frac{180\mu}{\delta^2}\frac{(1-f_l)^2}{f_l^3+B}u_i + \rho_{ref}g_i\beta\left(T-T_{ref}\right) \quad (2)$$

$$\frac{\partial \rho \mathrm{h}}{\partial t}+\frac{\partial \rho \mathrm{u_i h}}{\partial \mathrm{x_i}}=\frac{\partial}{\partial \mathrm{x_i}}\left(\mathrm{k}\frac{\partial \mathrm{T}}{\partial \mathrm{x_i}}\right) \quad (3)$$

where t is the time, u_i is the ith component of the velocity, μ is the viscosity, p is the pressure, h is the enthalpy, T is the temperature, ρ is the density, k is the thermal conductivity, and β is the thermal expansion coefficient. In this work, B is a small parameter with a value of 10^{-3} to avoid division by zero and δ is the approximate primary dendritic spacing, which was set to 1μm. Additionally, enthalpy is related to temperature according to the following:

$$\rho \mathrm{h}=\int_0^{\mathrm{T}} \rho \mathrm{c_p dT}+\rho \mathrm{Lf_l} \quad (4)$$

where $\mathrm{c_p}$ is the specific heat capacity, L is the latent enthalpy of fusion, and $\mathrm{f_l}$ is the volume fraction of liquid phase.

The thermal boundary condition including the heat source model at the metal-gas interface is specified as:

$$\mathrm{q_{ener}}=\frac{2\mathrm{Q}\eta}{\pi \mathrm{r_b^2}}\exp\left(\frac{-2((\mathrm{x}-\mathrm{V_s t})^2+\mathrm{y}^2)}{\mathrm{r_b^2}}\right) - \mathrm{h_c(T-T_\infty)}-\sigma_s\varepsilon(\mathrm{T^4}-\mathrm{T_{ref}^4}) \quad (5)$$

where Q is the laser power, η is the absorptivity, r_b is the laser beam radius, $\mathrm{V_s}$ is the scanning speed, $\mathrm{h_c}$ is the convective heat transfer coefficient, $\mathrm{T_\infty}$ is the ambient temperature, σ_s is the Stefan-Boltzmann constant, ε is the emissivity, and T_{ref} is the reference temperature.

The momentum boundary condition at the liquid-gas interface is:

$$F_{L/G}=\gamma n\kappa+\nabla_S T\frac{d\gamma}{dT} \quad (6)$$

where γ is the surface tension coefficient, n is the outward pointing normal of the surface, and κ is the curvature of the surface. The thermophysical properties of Inconel 625 and approximated processing conditions used in the simulations are summarized in Table 1. Densities taken from [50] for ambient and liquidus temperatures are used for the solid and liquid densities, respectively. Additionally, viscosity at the liquidus temperature taken from [51] was assumed to maintain a constant value within the melt pool. Temperature-dependent polynomial functions were fitted to experimental measurements [52] of thermal conductivity and specific heat capacity for the solid phase.

Table 1: Thermophysical properties and processing conditions used for the thermal-fluid flow model

Physical Property	Value	Reference
Solid density ($\mathrm{kg \cdot m^3}$)	8440	[50]
Liquid density ($\mathrm{kg \cdot m^3}$)	7640	[50]
Solidus temperature (K)	1563	[51]
Liquidus temperature (K)	1623	[51]
Solid specific heat capacity ($\mathrm{J \cdot kg^{-1} \cdot K^{-1}}$)	$0.2437T + 338.39$	[52]
Liquid specific heat capacity ($\mathrm{J \cdot kg^{-1} \cdot K^{-1}}$)	709.25	[50]
Solid thermal conductivity ($\mathrm{W \cdot m^{-1} \cdot K^{-1}}$)	$0.01530T + 5.2366$	[52]
Liquid thermal conductivity ($\mathrm{W \cdot m^{-1} \cdot K^{-1}}$)	30.078	[50]
Latent heat of fusion ($\mathrm{KJ \cdot kg^{-1} \cdot K^{-1}}$)	29.0	[50]
Dynamic viscosity (Pa · s)	7×10^{-3}	[51]
Coefficient of Thermal expansion ($\mathrm{K^{-1}}$)	5×10^{-5}	[51]
Preheat temperature (K)	353	-
Laser spot radius (μm)	45	-

3.2. Data-driven modeling techniques

This section aims to introduce the data-driven modeling techniques used in this study, namely the polynomial regression (PR) and Kriging methods. In general, the PR model uses the training data to estimate the optimal parameters of the polynomial formulation. On the other hand, the Kriging method offers an interpolation approach from which the prediction is derived based on correlations to existing data. The mathematical formulation of the predictive model can be expressed as:

$$y(\tilde{x})=f(\tilde{x})+\varepsilon \quad (7)$$

where $\mathrm{y}(\tilde{\mathrm{x}})$ represents the exact solution for new point $\tilde{\mathrm{x}}$, $\mathrm{f}(\tilde{\mathrm{x}})$ is a hypothetical function derived statistically from data that produces the model estimate, ε is error, and $\tilde{\mathrm{x}}$ represents a set of input variables. For different modeling approaches, the composition of each of these elements could be different.

3.2.1. Polynomial regression (PR) method

Similar to linear regression, PR formulates the relationship between the input variables x and the outcome y with higher-order variation [53]. The efficiency and accuracy of PR make it popular in various engineering domains. The quadratic polynomial function can be presented as:

$$\hat{y}=\beta_0+\sum_{i=1}^{k}\beta_i x_i+\sum_{i=1}^{k}\beta_{ii}x_i^2+\sum_i\sum_j\beta_{ij}x_i x_j \quad (8)$$

where β_0, β_i, β_{ii}, and β_{ij} are regression coefficients, and k is the number of design variables.

Moges, Tesfaye; Yang, Zhuo; Jones, Kevontrez; Feng, Shaw C.; Witherell, Paul; Lu, Yan. "Hybrid Modeling Approach for Melt Pool Prediction in Laser Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

3.2.2. Kriging method

Unlike traditional parametric modeling methods that can derive specific formulations, the Kriging method predicts a result based on the spatial correlation between an estimated data point and existing data points [54,55]. The general form of the Kriging approach can be presented as:

$$Z_E = \bar{Z} + \sum_{i=1}^{n} \lambda_i (Z_i - \bar{Z}) \tag{9}$$

where $\bar{Z}$ represents the regional mean value of the response and λ_i is the distance-correlated weight value, which is determined by computing the spatial correlation.

To calculate the weight factor, λ_i, one should first compute the spatial correlation R between data points. The value of the spatial correlation can be derived from:

$$R(\theta, x_i, x_j) = \prod_{l=1}^{n} exp(-\theta_i (x_{i,l} - x_{j,l})^2) \tag{10}$$

where $x_{i,l}$ is the l^{th} component of the i^{th} vector x_i [56]. $R(\theta, x_i, x_j)$ depends on the location of points x_i and x_j and the correlation parameter, θ.

3.3. Hybrid modeling approach

This section discusses development of the hybrid model based on a combination of computational and experimental data. The hybrid model applies a polynomial regression method to construct the initial simulation-based surrogate model using computational data obtained from the aforementioned CFD model (Section 3.1.). The Kriging method is then applied to model the residual error between computational and experimental results, using an adaptive modeling method to iteratively update the model and reduce the predictive errors of the hybrid model. Figure 2 outlines our workflow for the hybrid AM model construction.

3.3.1. Workflow of hybrid modeling approach

The proposed method uses physics-based computational data from a CFD model to construct the initial surrogate model by applying the polynomial regression method. This model can generally represent trends, but may include significant error when compared to the experimental data. The next step is to model the residual error between computational and experiment results. An adaptive modeling method can iteratively update the Kriging model to reduce the predictive errors of the hybrid model [57,58,59]. The process adaptively adjusts the training and validation datasets to approach higher predictive accuracy. To evaluate the hybrid model performance, the Average Relative Error Magnitude (AREM) is deployed for individual and global validation [58]. The formulation of AREM can be expressed as:

$$AREM = \frac{1}{m}\left(\frac{\sum_{i=1}^{m}|y_i - \hat{y}_i|}{y_i}\right) \quad (y_i \neq 0) \tag{11}$$

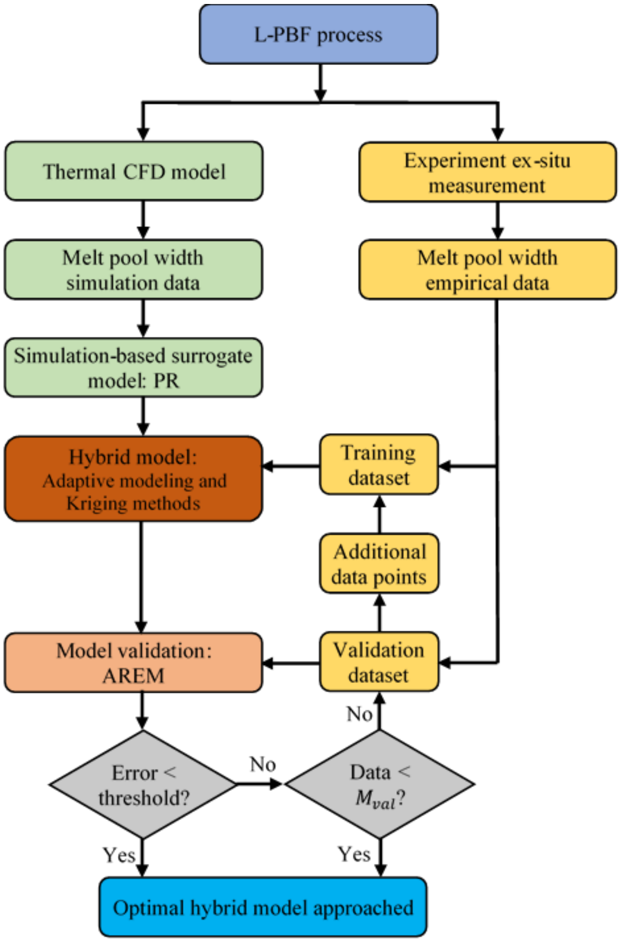

L-PBF process
Thermal CFD model
Experiment ex-situ measurement
Melt pool width simulation data
Melt pool width empirical data
Simulation-based surrogate model: PR
Hybrid model: Adaptive modeling and Kriging methods
Training dataset
Additional data points
Model validation: AREM
Validation dataset
No
Error < threshold?
No
Data < M_{val}?
Yes
Yes
Optimal hybrid model approached

Figure 2. Workflow for constructing AM hybrid model

where y_i is the observed value from given data, $\hat{y}$ is the value predicted by the surrogate model of the data points that were not selected to construct the surrogate model, and m is the number of data points.

3.3.2. Algorithm for hybrid model development

This section provides the algorithmic approach to build the proposed AM hybrid model. We first generate N_{sim} simulation data to construct the initial surrogate model. Similarly, M_{meas} experimental data are measured to improve and validate the hybrid model. The adaptive Kriging model starts with M_{in} initial experimental data points and M_{up} additional points for updating. To validate the final hybrid model, M_{val} experimental data points are used. Therefore, the training-validation data ratio is $(M_{in} + M_{up}) : M_{val}$. The case study in Section 4 uses a total of 72 simulation data points and 21 experimental data points. Here, it is important to note that, in the limited number of experimental data, the predetermined number of validation data can be reached before the error value is less than the threshold. In this case, all

Moges, Tesfaye; Yang, Zhuo; Jones, Kevontrez; Feng, Shaw C.; Witherell, Paul; Lu, Yan. "Hybrid Modeling Approach for Melt Pool Prediction in Laser Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

the remaining validation datapoints will be used to validate the hybrid model. This situation can be avoided by providing a greater number of experimental data.

Algorithm for hybrid modeling approach

```
Step 1: Generate experimental and computational data
Step 2: Construct initial surrogate model
    Step 2.1: Construct initial surrogate model using $N_{sim}$ by
              PR method
    Step 2.2: Set target threshold for error comparison
    Step 2.3: Validate the surrogate model using additional
              simulation data
              if error > threshold   do
                 Back to Step 1 and generate more simulation
                   data
              else
                 Simulation-based surrogate model is ready and
                 proceed to Step 3
              end if
Step 3: Prepare the training and validation datasets of $M_{meas}$
        experimental data
    Step 3.1: Compare the results of the surrogate model
              against $M_{meas}$ experimental data
    Step 3.2: Select $M_{in}$ experimental data points with largest
              error to construct the initial training dataset
    Step 3.3: Store the rest of the data that would be used as
              validation dataset
Step 4: Construct the hybrid model
    Step 4.1: Calculate the residual error between results from
              surrogate model and experimental training
              dataset
    Step 4.2: Build the surrogate model for residual error by
              Kriging method
    Step 4.3: Combine the PR and Kriging methods to
              construct the hybrid model
Step 5: Hybrid model validation
    Step 5.1: Use the validation dataset to validate the hybrid
              model
    Step 5.2: Set target threshold for error comparison
              if error > threshold do
                 if data > $M_{val}$ do
                       Add an $M_{up}$ additional data point with
                       largest error to the training dataset
                 end if
              else
                 Validate the hybrid model using validation
                 dataset and proceed to Step 6
              end if
    Step 5.3: Eliminate the selected data point from the $M_{val}$
              validation dataset
    Step 5.4: Go back to Step 4 to validate the hybrid model
Step 6: Approach the final hybrid model
```

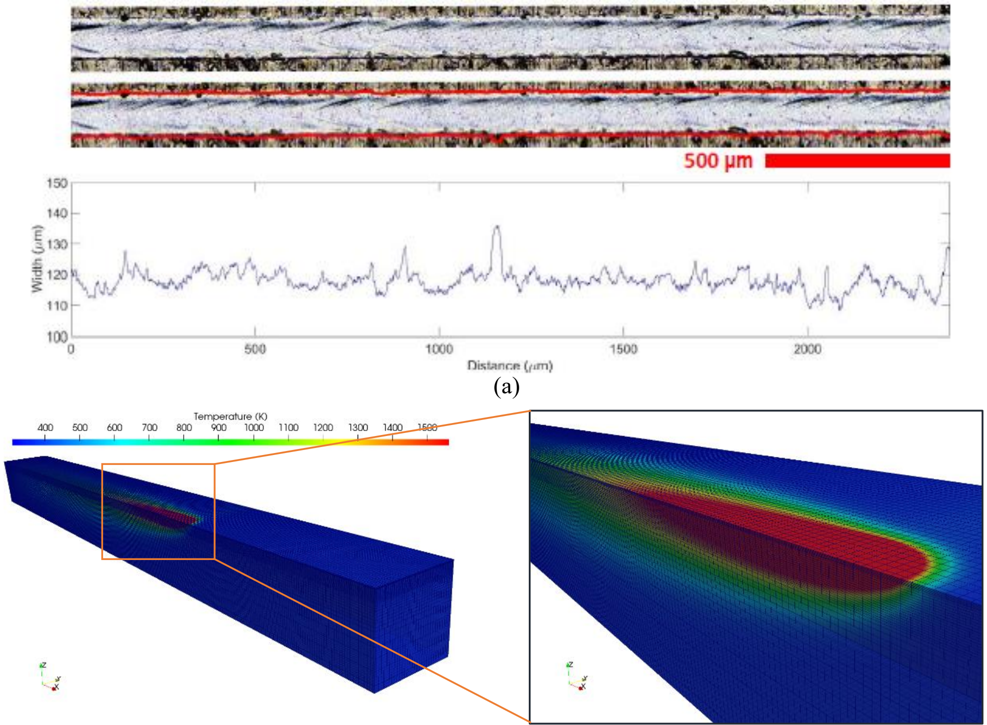
4. CASE STUDY: HYBRID MODELING OF MELT POOL WIDTH

This section demonstrates the proposed hybrid model presented in Section 3.3 and discusses its performance by comparing it against computational and experimental data. As previously stated, melt pool widths obtained from CFD model and measurement data are used as output QoIs to demonstrate the hybrid model's capabilities. The primary input variables used for the simulations and experiments are laser power and scan speed. The range of laser power used for the simulations and experiments is from 49W – 285W and 100W – 195W, respectively. Whereas, the scan speed used for both the simulations and experiments ranges from 100mm/s – 1400mm/s.

4.1. Data for hybrid model

The experiments were performed on Inconel 625 bare plates and ex-situ measurements were conducted to record melt pool widths using optical microscope [60]. The melt pool widths were measured from the optical image of a 1mm long scan track by manually tracing the edges of the track and averaging the distance between the traces at different locations as shown in Figure 3(a). The average and standard deviation of the measured melt pool widths at different combinations of laser power and scan speed are given in Table 2. Similarly, the thermal CFD model is simulated as shown in Figure 3(b), and melt pool widths are extracted for a given input variables. The simulation results for the corresponding experimental data along with the relative percentage error are also given in Table 2.

The average relative percentage error of the simulations against measurement results is 20.78%. This highlights that the physics-based computational model by itself induces major discrepancy against the experimental data. This discrepancy may be due to the different assumptions taken during model formulation including powder particle distribution, spattering of molten metal, gas-liquid-solid interaction, mass loss due to chemical reactions, and others. This discrepancy is due to model uncertainty and also the uncertainty associated with measurements, including sensor error and imprecise measurement methods [15,22].



(a)

(b)

Figure 3. Experimental-based melt pool width from optical image [60] (a) and melt pool region from the CFD simulation model (b)

Table 2. Melt pool widths: measurement, simulation, and % relative error

No. of Run	Laser power (W)	Scan speed (mm/s)	Melt pool width (μm)		% relative error
			Measurement with St. dev	Simu-lation	
1	100	100	259.77 (8.22)	175	32.63
2	100	200	177.26 (4.73)	146	17.64
3	100	400	123.39 (7.73)	120	2.75
4	100	600	86.79 (3.43)	110	26.74
5	100	1000	73.918 (2.47)	94.1	27.30
6	100	1200	72.123 (2.87)	88.2	22.29
7	100	1400	68.932 (2.19)	90	30.56
8	150	200	225.87 (11.16)	161	28.72
9	150	400	166.96 (9.90)	132	20.94
10	150	600	146.89 (8.70)	121	17.63
11	150	800	105.73 (5.15)	110	4.04
12	150	1000	86.922 (3.26)	110	26.55
13	150	1200	93.365 (4.62)	105	12.46
14	150	1400	85.735 (5.34)	101	17.80
15	195	100	362.25 (13.33)	206	43.13
16	195	200	256.05 (16.98)	167	34.78
17	195	400	188.7 (10.69)	142	24.75
18	195	600	149.62 (5.28)	130	13.11
19	195	800	126.3 (4.33)	120	4.99
20	195	1000	107.17 (4.55)	113	5.44
21	195	1200	90.15 (4.97)	110	22.02

4.2. Implementation of the proposed hybrid model

Figure 4 shows the schematic framework for the case study implementation of the hybrid model for melt pool width prediction. First, the CFD model is simulated and 72 data points of melt pool widths were extracted for the given combinations of laser power and scan speed, as shown in PV map (left) in Figure 4. Similarly, experimental data provided 21 data points of melt pool widths are measured for the given combinations of laser power and scan speed, as shown in the PV map (right). All simulation data are used to construct the initial surrogate model using a polynomial regression method. The adaptive Kriging model starts with 5 initial data points and 10 additional points for updating. As a result, 6 experimental data points are excluded from the model construction and are used to validate the model. The training-validation data ratio is 15:6. The hybrid model is then used to predict melt pool widths for other possible combinations of laser power and scan speed in the design space with good accuracy and computational time.

The accuracy of the proposed hybrid model mainly depends on the accuracies and the number of computational and experimental data used and the data-driven techniques applied. Obtaining more experimental data used for training, updating, and validating the proposed hybrid model is crucial for improving the accuracy.

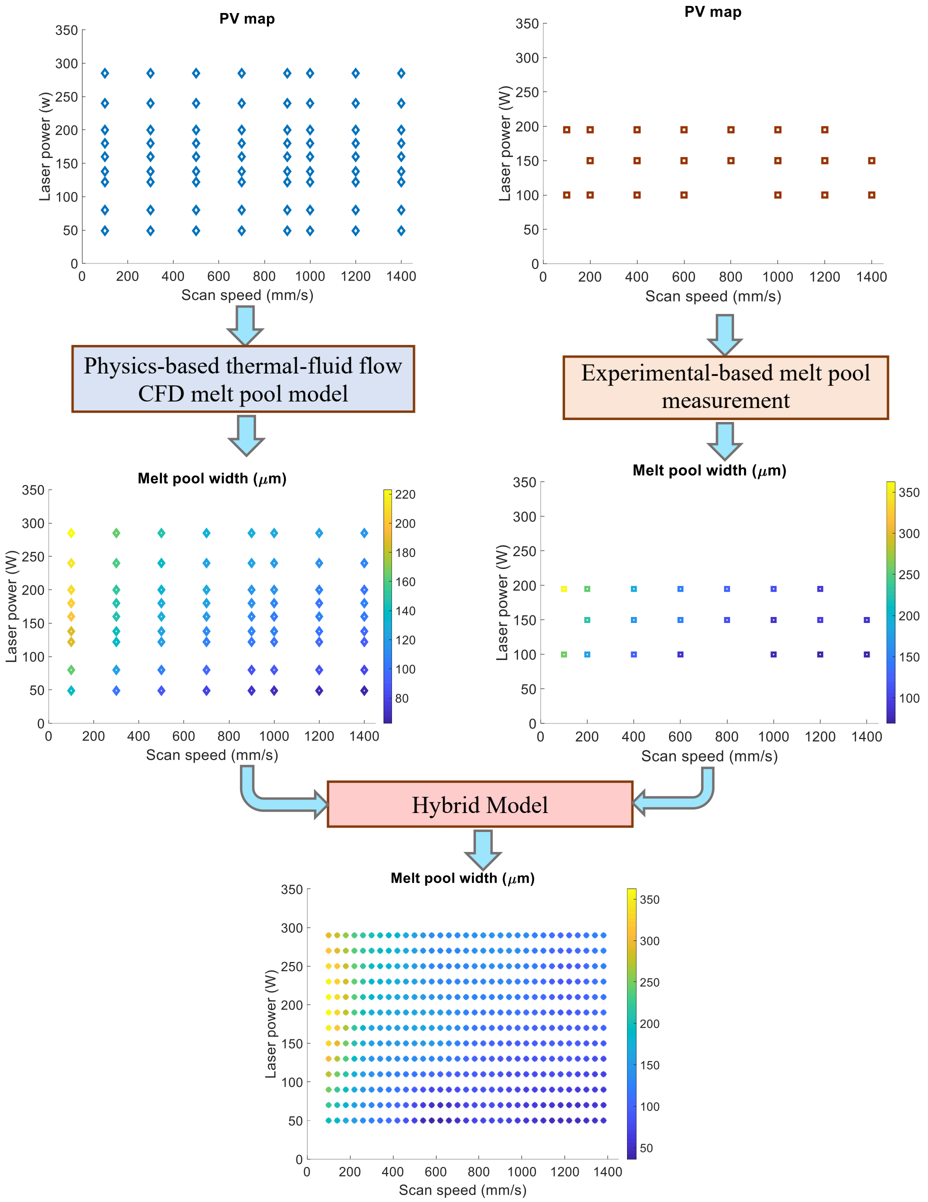
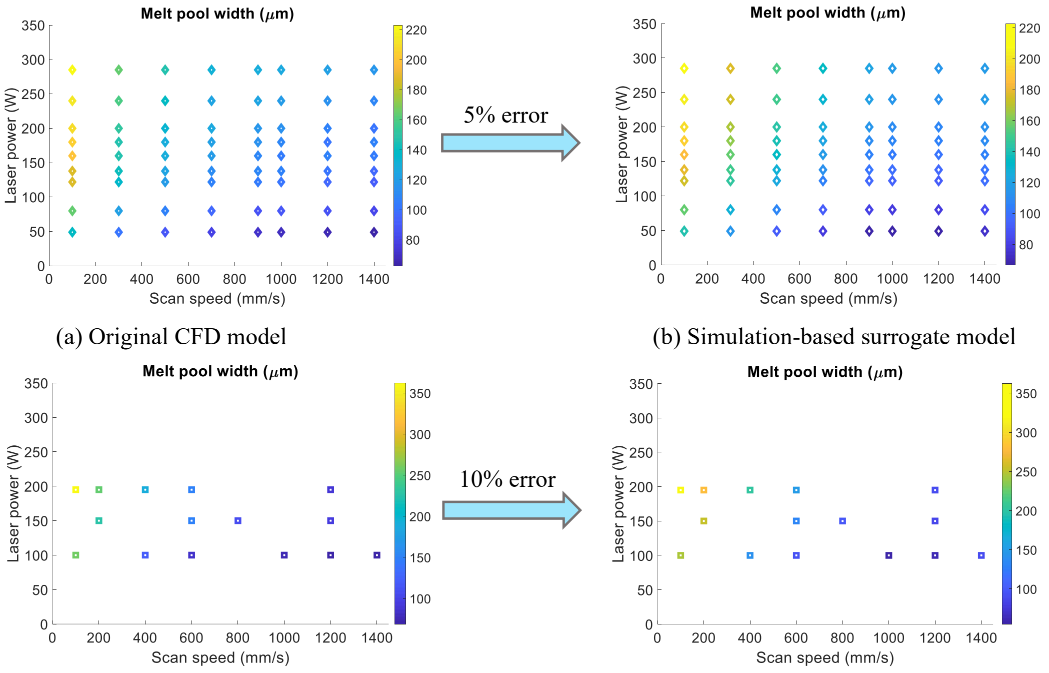


Figure 4 Schematic implementation framework of hybrid model of melt pool width

4.3. Verification of Surrogate models

To develop and evaluate the hybrid model, both computational-based and experimental-based surrogate models are constructed using corresponding data. To evaluate the accuracy of these surrogate models and compare them with the original simulation and measured data, model verification is conducted for the same processing parameters. The computational-based surrogate model is compared to the original CFD model as shown in Figure 5(a) and (b), and the average error is about 5%. The cause of this discrepancy may be due to the variabilities in data-driven modeling approaches. Similarly, the experimental-based surrogate model has around 10% difference when compared to the original measurement data as shown in Figure 5(c) and (c). This error is mainly attributed to the limited number of measurement data points (only 15) used to build the surrogate model.



(a) Original CFD model

(b) Simulation-based surrogate model

(c) Original measurement data

(d) Experimental-based surrogate model

Figure 5. Comparison of original simulation model and measurement data with surrogate models

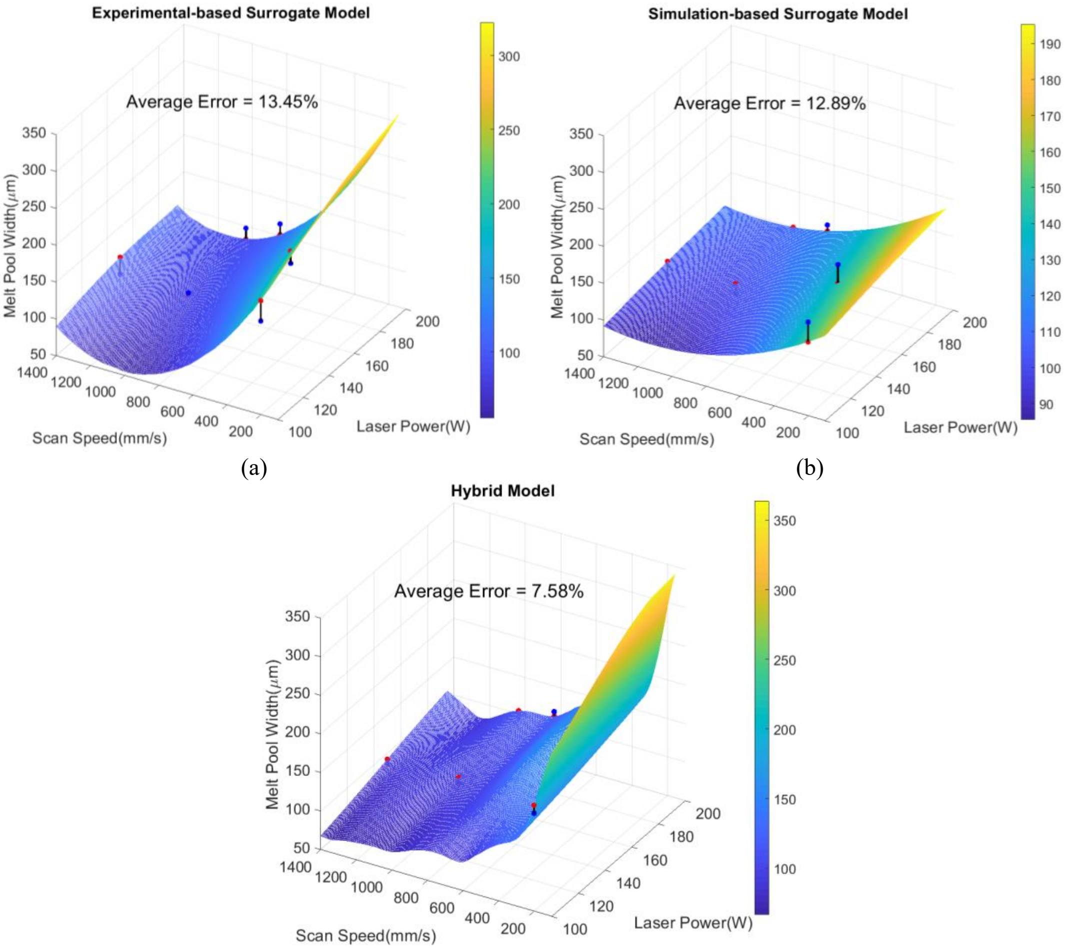
4.4. Performance evaluation of hybrid model

In order to evaluate the performance of the proposed model, we compared the average relative error of hybrid model to the experimental-based surrogate model and computational-based surrogate model as shown in Figure 6 and Table 3. The blue and red dots in the graph represent the measured and predicted results with the validation parameters, respectively.

The results of melt pool widths predicted at the 6 validation data points (that are not used in the model construction) using the experimental-based surrogate model has an average relative error of 13.45%, as shown in Figure 6(a). Similarly, the melt pool widths predicted using the simulation-based surrogate model has an average error of 12.89% as shown in Figure 6(b). Figure 6(c) depicts the prediction of melt pool widths using the proposed hybrid model. The hybrid model has an average error of 7.58%. Table 3 provides the input parameters used for validation, the measured results, simulated results, predictive results from experimental-based and simulation-based surrogate models, and their deviations from the measured results using percentage relative error in parentheses. It can be observed that the hybrid model predicted melt pool width with better accuracy than both simulation-based and experimental-based surrogate models.

Table 3. Comparing the proposed hybrid model to the experimental validation data points

No. of Run	Laser power (W)	Scan speed (mm/s)	Measurement results	CFD simulation results	Experiment-based surrogate model	Simulation-based surrogate model	Hybrid model
1	100	200	177.26	146 (17.64%)	204.92 (15.60%)	150.24 (15.24%)	187.71 (5.90%)
2	150	400	166.96	132 (20.94%)	183.63 (9.98%)	141.91 (15.00%)	148.26 (11.20%)
3	150	1000	86.922	110 (26.55%)	84.13 (3.21%)	100.94 (16.13%)	95.21 (9.54%)
4	150	1400	85.735	101 (17.80%)	108.64 (26.72%)	104.52 (21.92%)	92.20 (7.54%)
5	195	800	126.3	120 (4.99%)	111.82 (11.46%)	118.23 (6.39%)	120.34 (4.72%)
6	195	1000	107.17	110 (2.64%)	92.53 (13.66%)	110.05 (2.69%)	114.23 (6.59%)
	Average % relative error			**(15.09%)**	**(13.45%)**	**(12.89%)**	**(7.58%)**



(a)

(b)

(c)

Figure 6. Melt pool width prediction using (a) experimental-based surrogate model, (b) simulation-based surrogate model, (c) proposed hybrid model

Therefore, by integrating the physics-based and experimentally obtained data using the proposed hybrid modeling approach, melt pool widths can be predicted with better accuracy. It can be concluded that with a few experimental data, the overall prediction accuracy of physics-based numerical models can be improved through the proposed approach. Since the hybrid model is a data-driven approach, for new sets of input parameters, the melt pool widths can be predicted in a few seconds. Thus, the hybrid model is computationally efficient compared to pure physics-based models. Due to improved accuracy and computational efficient, the proposed hybrid model better suited for real-time monitoring, control and optimization.

5. FUTURE WORK: AM ONTOLOGY FOR MODEL SELECTION

Our case study demonstrated the ability to create a hybrid model for a specific application, with comparable parameter sets. However, the context under which models are run and experiments are conducted is not always as straight forward. To help resolve potential discrepancies in the data sources, we propose there is a role for ontologies. Ontology can be used to capture the rapidly evolving knowledge of the AM process, computational models, uncertainty sources, and design for AM in an organized structure to help users to interoperate and reuse information [47,49,61]. As stated previously, there are numerous physics-based models ranging from low-fidelity to medium-fidelity to high-fidelity. A well-founded AM ontology can be used to capture the complex interconnections between these models as well as data-driven models and provide useful information for model composition towards developing a more accurate and reliable predictive metamodel [62,63]. In this section, we introduce the conceptual use of AM ontology for providing structured information of the different AM models to help develop a more accurate hybrid model.

An ontology consists of various concepts and correlations among entities. The AM ontology offered in Moges et al. [47]

captures the different features of AM models including assumptions, considered and neglected phenomena, model inputs and outputs, and uncertainty sources. The ontology attempted to capture these features at the five stages of the process: powder layer formation, laser-heat source interaction, melt pool formation, solidification, and residual stress formation. Different physics-based models have been developed to simulate each of these stages of the process. For instance, to simulate the melt pool behavior in L-PBF, there are Rosenthal-based analytical models, FEM thermal models, path-level thermal simulation model [64], CFD models, and Lattice Boltzmann method (LBM) models [14]. Similarly, to determine the amount of absorbed energy by the powder particles, there are different methods including radiation transfer, ray tracing, and Beer-Lambert approach [14]. Since these models are developed based on different methods and formulations, they have different pros and cons in terms of prediction accuracy and computational time. Even some models predict more accurately at a specific region of a PV map than the other regions.

The hybrid model developed in this paper uses computationally predicted data obtained from one physics-based model namely CFD model. It was shown in Section 4 that this hybrid model improves the predictive accuracy of the melt pool width compared to the original physics-based model. However, in order to further improve the accuracy, instead of using simulation data obtained from a single physics-based model, a hybrid model that uses different simulation data obtained from different physics-based models could be more reliable. For this, a systematic approach that leverages the AM ontology is needed to select the simulation data obtained from different physics-based models and integrate them with experimental data to build a more precise hybrid model.

In the future, a more reliable approach need to be developed for (a) investigating the capability of different physics-based models and data-driven techniques, (b) developing strategic approaches to select the different simulation data, (c) integrating the different simulation data with the experimental data using various data-driven methods. Hence, ontology can be an essential tool for strategically selecting models based on their inherent key features. Furthermore, incorporating different data-driven techniques into the existing AM ontology will enable us to select and apply a more suitable technique for developing a hybrid model to predict other output QoIs.

6. CONCLUSION

In this study we proposed a methodology to integrate physics-based data and experimentally measured data into a hybrid model that enables fast and accurate predictions in the L-PBF process. We proposed a hybrid model which comprised of data generated from both numerically predicted and experimentally recorded melt pool widths for various combinations of laser power and scan speed. The numerical results were obtained from a thermal CFD model and ex-situ cross-sectioning was used to gather the experimental results.

In our hybrid modeling approach, we first constructed an initial simulation-based surrogate model using simulation data by applying the polynomial regression method. Then, we applied the Kriging method to model the residual error between the simulated and experimental results using an adaptive modeling method to reduce predictive errors of the hybrid model. The performance of the proposed model is evaluated by comparing predicted results of melt pool widths from a CFD model, a simulation-based surrogate model, an experimental-based surrogate model, and the hybrid model against experimentally measured data. The results showed that on average, the hybrid model had the highest accuracy out of all the models. Additionally, due to improved accuracy and computational efficiency, the proposed hybrid model better suited for real-time process control.

In order to further improve the accuracy of the hybrid model, instead of using data obtained from a single physics-based model, integrating data obtained from multiple physics-based models with experimentally measured data could be more reliable. To address this, AM ontology can be used as an essential tool to help understand the capability of the different physics-based models and data-driven techniques and select the most accurate model at a specific region in the PV map. In the future, the knowledge captured in AM ontology can be leveraged for developing a more accurate hybrid modeling approach by providing predictive capabilities of multiple models. We view this work as a step towards effectively and efficiently utilizing physics-based models, data-driven approaches, experimental-based measurement data, and AM ontology to build reliable and robust predictive models that can be applied for L-PBF process control.

Acknowledgment

The authors gratefully acknowledge Jason Fox from the Engineering Laboratory at National Institute of Standards and Technology (NIST) for conducting experiments and providing measurement data.

DISCLAIMER

Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by National Institute of Standards and Technology (NIST); nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

References

[1] Bourell, D. L., Beaman, J. J., Marcus, H. L., and Barlow, J. W., 1990, "Solid Freefor Fabrication: An Advanced Manufacturing Approach," Proc. Annu. Int. Solid Free. Fabr. Symp., pp. 1–7.

[2] Petrovic, V., Vicente Haro Gonzalez, J., Jordá Ferrando, O., Delgado Gordillo, J., Ramón Blasco Puchades, J., and Portolés Griñan, L., 2011, "Additive Layered Manufacturing: Sectors of Industrial Application Shown

through Case Studies," Int. J. Prod. Res., **49**(4), pp. 1061–1079.

[3] Yan, C., Hao, L., Hussein, A., and Raymont, D., 2012, "Evaluations of Cellular Lattice Structures Manufactured Using Selective Laser Melting," Int. J. Mach. Tools Manuf., **62**, pp. 32–38.

[4] Guo, N., and Leu, M. C., 2013, "Additive Manufacturing: Technology, Applications and Research Needs," Front. Mech. Eng., **8**(3), pp. 215–243.

[5] Herderick, E., 2011, "Additive Manufacturing of Metals: A Review," Mater. Sci. Technol. Conf. Exhib. 2011, MS T'11, **2**(176252), pp. 1413–1425.

[6] Hofmann, D. C., Roberts, S., Otis, R., Kolodziejska, J., Dillon, R. P., Suh, J. O., Shapiro, A. A., Liu, Z. K., and Borgonia, J. P., 2014, "Developing Gradient Metal Alloys through Radial Deposition Additive Manufacturing," Sci. Rep., **4**.

[7] Gu, D., 2015, *Laser Additive Manufacturing (AM): Classification, Processing Philosophy, and Metallurgical Mechanisms*, Springer Berlin Heidelberg.

[8] Bourell, D. L., Leu, M. C., and Rosen, D. W., 2009, "Roadmap for Additive Manufacturing: Identifying the Future of Freeform Processing," Rapid Prototyp. J., **5**(4), pp. 169–178.

[9] Read, N., Wang, W., Essa, K., and Attallah, M. M., 2015, "Selective Laser Melting of AlSi10Mg Alloy: Process Optimisation and Mechanical Properties Development," Mater. Des., **65**, pp. 417–424.

[10] Criales, L. E., Arısoy, Y. M., and Özel, T., 2016, "Sensitivity Analysis of Material and Process Parameters in Finite Element Modeling of Selective Laser Melting of Inconel 625," Int. J. Adv. Manuf. Technol., **86**(9–12), pp. 2653–2666.

[11] Criales, L. E., Arısoy, Y. M., Lane, B., Moylan, S., Donmez, A., and Özel, T., 2017, "Laser Powder Bed Fusion of Nickel Alloy 625: Experimental Investigations of Effects of Process Parameters on Melt Pool Size and Shape with Spatter Analysis," Int. J. Mach. Tools Manuf., **121**(March), pp. 22–36.

[12] Schoinochoritis, B., Chantzis, D., and Salonitis, K., 2014, "Simulation of Metallic Powder Bed Additive Manufacturing Processes with the Finite Element Method: A Critical Review," Proc. Inst. Mech. Eng. Part B J. Eng. Manuf., **231**(1), pp. 96–117.

[13] King, W. E., Anderson, A. T., Ferencz, R. M., Hodge, N. E., Kamath, C., Khairallah, S. A., and Rubenchik, A. M., 2015, "Laser Powder Bed Fusion Additive Manufacturing of Metals; Physics, Computational, and Materials Challenges," Appl. Phys. Rev., **2**(4), p. 041304.

[14] Moges, T., Ameta, G., and Witherell, P., 2019, "A Review of Model Inaccuracy and Parameter Uncertainty in Laser Powder Bed Fusion Models and Simulations," J. Manuf. Sci. Eng., **141**(4), p. 040801.

[15] Moges, T., Yan, W., Lin, S., Ameta, G., Fox, J., and Witherell, P., 2018, "Quantifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models and Simulations," *Solid Freeform Fabrication Symposium*, pp. 1913–1928.

[16] Wu, Q., Lu, J., Liu, C., Fan, H., Shi, X., Fu, J., and Ma, S., 2017, "Effect of Molten Pool Size on Microstructure and Tensile Properties of Wire Arc Additive Manufacturing of Ti-6Al-4V Alloy," Materials (Basel)., **10**(7), pp. 1–11.

[17] Malekipour, E., and El-Mounayri, H., 2018, "Common Defects and Contributing Parameters in Powder Bed Fusion AM Process and Their Classification for Online Monitoring and Control: A Review," Int. J. Adv. Manuf. Technol., **95**(1–4), pp. 527–550.

[18] Marrey, M., Malekipour, E., El-Mounayri, H., and Faierson, E. J., 2019, "A Framework for Optimizing Process Parameters in Powder Bed Fusion (PBF) Process Using Artificial Neural Network (ANN)," Procedia Manuf., **34**, pp. 505–515.

[19] Hu, Z., and Mahadevan, S., 2017, "Uncertainty Quantification and Management in Additive Manufacturing: Current Status, Needs, and Opportunities," Int. J. Adv. Manuf. Technol., **93**, pp. 2855–2874.

[20] Smith, J., Xiong, W., Yan, W., Lin, S., Cheng, P., Kafka, O. L., Wagner, G. J., Cao, J., and Liu, W. K., 2016, "Linking Process, Structure, Property, and Performance for Metal-Based Additive Manufacturing: Computational Approaches with Experimental Support," Comput. Mech., **57**(4), pp. 583–610.

[21] Devesse, W., De Baere, D., and Guillaume, P., 2014, "The Isotherm Migration Method in Spherical Coordinates with a Moving Heat Source," Int. J. Heat Mass Transf., **75**, pp. 726–735.

[22] Lopez, F., Witherell, P., and Lane, B., 2016, "Identifying Uncertainty in Laser Powder Bed Fusion Additive Manufacturing Models," J. Mech. Des., **138**(November), pp. 1–4.

[23] Lin, S., Smith, J., Liu, W. K., and Wagner, G. J., 2017, "An Energetically Consistent Concurrent Multiscale Method for Heterogeneous Heat Transfer and Phase Transition Applications," Comput. Methods Appl. Mech. Eng., **315**, pp. 100–120.

[24] Wolff, S. J., Lin, S., Faierson, E. J., Liu, W. K., Wagner, G. J., and Cao, J., 2017, "A Framework to Link Localized Cooling and Properties of Directed Energy Deposition (DED)-Processed Ti-6Al-4V," Acta Mater., **132**, pp. 106–117.

[25] Romano, J., Ladani, L., and Sadowski, M., 2015, "Thermal Modeling of Laser Based Additive Manufacturing Processes within Common Materials," Procedia Manuf., **1**, pp. 238–250.

[26] Li, Y., and Gu, D., 2014, "Parametric Analysis of Thermal Behavior during Selective Laser Melting Additive Manufacturing of Aluminum Alloy Powder," Mater. Des., **63**, pp. 856–867.

[27] Manvatkar, V., De, A., and Debroy, T., 2014, "Heat

Transfer and Material Flow during Laser Assisted Multi-Layer Additive Manufacturing," J. Appl. Phys., **116**(12), pp. 1–8.

[28] Gan, Z., Lian, Y., Lin, S. E., Jones, K. K., Liu, W. K., and Wagner, G. J., 2019, "Benchmark Study of Thermal Behavior, Surface Topography, and Dendritic Microstructure in Selective Laser Melting of Inconel 625," Integr. Mater. Manuf. Innov., **8**(2), pp. 178–193.

[29] Gan, Z., Liu, H., Li, S., He, X., and Yu, G., 2017, "Modeling of Thermal Behavior and Mass Transport in Multi-Layer Laser Additive Manufacturing of Ni-Based Alloy on Cast Iron," Int. J. Heat Mass Transf., **111**, pp. 709–722.

[30] Mukherjee, T., Wei, H. L., De, A., and DebRoy, T., 2018, "Heat and Fluid Flow in Additive Manufacturing—Part I: Modeling of Powder Bed Fusion," Comput. Mater. Sci., **150**(February), pp. 304–313.

[31] Mukherjee, T., and DebRoy, T., 2018, "Mitigation of Lack of Fusion Defects in Powder Bed Fusion Additive Manufacturing," J. Manuf. Process., **36**(November), pp. 442–449.

[32] Khairallah, S. A., Anderson, A. T., Rubenchik, A., and King, W. E., 2016, "Laser Powder-Bed Fusion Additive Manufacturing: Physics of Complex Melt Flow and Formation Mechanisms of Pores, Spatter, and Denudation Zones," Acta Mater., **108**, pp. 36–45.

[33] Lee, Y., and Farson, D. F., 2016, "Simulation of Transport Phenomena and Melt Pool Shape for Multiple Layer Additive Manufacturing," J. Laser Appl., **28**(1), p. 012006.

[34] Wen, S. Y., Shin, Y. C., Murthy, J. Y., and Sojka, P. E., 2009, "Modeling of Coaxial Powder Flow for the Laser Direct Deposition Process," Int. J. Heat Mass Transf., **52**(25–26), pp. 5867–5877.

[35] Ghosh, S., Ma, L., Levine, L. E., Ricker, R. E., Stoudt, M. R., Heigel, J. C., and Guyer, J. E., 2018, "Single-Track Melt-Pool Measurements and Microstructures in Inconel 625," Jom, **70**(6), pp. 1011–1016.

[36] Tapia, G., Khairallah, S., Matthews, M., King, W. E., and Elwany, A., 2018, "Gaussian Process-Based Surrogate Modeling Framework for Process Planning in Laser Powder-Bed Fusion Additive Manufacturing of 316L Stainless Steel," Int. J. Adv. Manuf. Technol., **94**(9–12), pp. 3591–3603.

[37] Yang, Z., Yan, L., Yeung, H., and Krishnamurty, S., 2019, "From Scan Strategy to Melt Pool Prediction:A Neighboring-Effect Modeling Method," *Proceedings of the ASME Design Engineering Technical Conference*, pp. 1–11.

[38] Wang, Z., Liu, P., Ji, Y., Mahadevan, S., Horstemeyer, M. F., Hu, Z., Chen, L., and Chen, L. Q., 2019, "Uncertainty Quantification in Metallic Additive Manufacturing Through Physics-Informed Data-Driven Modeling," Jom, **71**(8), pp. 2625–2634.

[39] Razvi, S. S., Feng, S., Narayanan, A., Lee, Y.-T. T., and Witherell, P., 2019, "A Review of Machine Learning Applications in Additive Manufacturing," *Volume 1: 39th Computers and Information in Engineering Conference*, American Society of Mechanical Engineers, Anaheim, CA, USA.

[40] Fathi, A., and Mozaffari, A., 2014, "Vector Optimization of Laser Solid Freeform Fabrication System Using a Hierarchical Mutable Smart Bee-Fuzzy Inference System and Hybrid NSGA-II/Self-Organizing Map," J. Intell. Manuf., **25**(4), pp. 775–795.

[41] Lu, Z. L., Li, D. C., Lu, B. H., Zhang, A. F., Zhu, G. X., and Pi, G., 2010, "The Prediction of the Building Precision in the Laser Engineered Net Shaping Process Using Advanced Networks," Opt. Lasers Eng., **48**(5), pp. 519–525.

[42] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Witherell, P. W., and Lopez, F., 2018, "Dynamic Metamodeling for Predictive Analytics in Advanced Manufacturing," Smart Sustain. Manuf. Syst., **2**(1), p. 20170013.

[43] Kamath, C., 2016, "Data Mining and Statistical Inference in Selective Laser Melting," Int. J. Adv. Manuf. Technol., **86**(5–8), pp. 1659–1677.

[44] Tran, H. C., and Lo, Y., 2019, "Systematic Approach for Optimal Determining Optimal Processing Parameters to Produce Part with High Density in Selective Laser Melting Process," Int. J. Adv. Manuf. Technol.

[45] Solomatine, D., Abrahart, R., and See, L., 2008, "Data-Driven Modelling: Concepts, Approaches and Experiences," *Practical Hydroinformatics*, pp. 17–30.

[46] Reinhart, R. F., Shareef, Z., and Steil, J. J., 2017, "Hybrid Analytical and Data-Driven Modeling for Feed-Forward Robot Control," Sensors (Switzerland), **17**(2), pp. 1–19.

[47] Moges, T., Witherell, P., and Ameta, G., 2019, "ON CHARACTERIZING UNCERTAINTY SOURCES IN LASER POWDER BED FUSION ADDITIVE MANUFACTURING MODELS," *Proceedings of the ASME 2019 International Mechanical Engineering Congress and Exposition IMECE2019 November 11-14, 2019, Salt Lake City, UT, USA*, pp. 1–15.

[48] Witherell, P., Feng, S. C., Martukanitz, R., Simpson, T. W., John, D. B. S., Michaleris, P., Liu, Z. K., and Chen, L. Q., 2014, "Toward Metamodels for Composable and Reusable Additive Manufacturing Process Models," Proc. ASME Des. Eng. Tech. Conf., **1A**, pp. 1–10.

[49] Assouroko, Ibrahim; Lopez, Felipe; Witherell, P., 2016, "A Method for Characterizing Model Fidelity in Laser Powder Bed Fusion Additive Manufacturing," *Proceedings of the ASME 2016 International Mechanical Engineering Congress & Exposition ASME IMECE 2016 November 11-17, 2016, Phoenix, Arizona, USA*, pp. 1–13.

[50] Capriccioli, A., and Frosi, P., 2009, "Multipurpose ANSYS FE Procedure for Welding Processes Simulation," Fusion Eng. Des., **84**(2–6), pp. 546–553.

[51] Pawel, R. E., and Williams, R. K., 1985, "Survey of

Physical Property Data for Several Alloys," Oak Ridge Natl. Lab., (ORNL/TM-9616).

[52] Corporation, S. M., 2013, "Inconel Alloy 625," www.Specialmetals.com, **625**(2), pp. 1–28.

[53] Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F., 1998, "Comparison of Response Surface and Kriging Models for Multidisciplinary Design Optimization," *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, p. 4755.

[54] Cressie, N., 2015, *Statistics for Spatial Data*, John Wiley & Sons, Inc.

[55] Simpson, T. W., Booker, A. J., Ghosh, D., Giunta, A. A., Koch, P. N., and Yang, R., 2004, "Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion," Struct. Multidiscip. Optim., **27**(5), pp. 302–313.

[56] Sacks, J., Welch, W., Mitchell, T., and Wynn, H., 1989, "Design and Analysis of Computer Experiments," Stat. Sci., **4**(4), pp. 409–423.

[57] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Lu, Y., and Witherell, P., 2017, "Investigating Grey-Box Modeling for Predictive Analytics in Smart Manufacturing," Proc. ASME Des. Eng. Tech. Conf., **2**B-**2017**, pp. 1–10.

[58] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., and Lopez, F., 2016, "Investigating Predictive Metamodeling for Additive Manufacturing," Proc. ASME Des. Eng. Tech. Conf., **1**A-**2016**, pp. 1–10.

[59] Shao, T., and Krishnamurty, S., 2008, "A Clustering-Based Surrogate Model Updating Approach to Simulation-Based Engineering Design," J. Mech. Des. Trans. ASME, **130**(4), pp. 1–13.

[60] Fox, J. C., Lane, B. M., and Yeung, H., 2017, "Measurement of Process Dynamics through Coaxially Aligned High Speed Near-Infrared Imaging in Laser Powder Bed Fusion Additive Manufacturing," Proc. SPIE 10214, Thermosense Therm. Infrared Appl. XXXIX, **1**(301), p. 1021407.

[61] Kim, S., Rosen, D. W., Witherell, P., and Ko, H., 2019, "A Design for Additive Manufacturing Ontology to Support Manufacturability Analysis," J. Comput. Inf. Sci. Eng., **19**(December), pp. 041014-1–10.

[62] Roh, B., Kumara, S. R. T., Simpson, T. W., and Witherell, P., 2016, "Ontology-Based Laser and Thermal Metamodels for Metal-Based Additive Manufacturing," Proc. ASME 2016 Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf. IDETC/CIE 2016 August 21-24, 2016, Charlotte, North Carolina, pp. 1–8.

[63] Witherell, P., Feng, S., Simpson, T. W., Saint John, D. B., Michaleris, P., Liu, Z.-K., Chen, L.-Q., and Martukanitz, R., 2014, "Toward Metamodels for Composable and Reusable Additive Manufacturing Process Models," J. Manuf. Sci. Eng., **136**(6), p. 061025.

[64] Zhang, Y., Shapiro, V., and Witherell, P., 2019, "TOWARDS THERMAL SIMULATION OF POWDER BED FUSION ON PATH LEVEL," *Proceedings of the ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2019 August 18-21, 2019, Anaheim, CA, USA*.

Utility Proportional Resource Allocation for Users with Diverse SLAs in Virtualized Radio Access Networks

Behnam Rouzbehani[1], Vladimir Marbukh[2], Kamran Sayrafian[2]

[1]IST-University of Lisbon/INESC-ID
Av. Rovisco Pais 1, 1049-001
Lisbon, Portugal

[2]National Institute of Standards & Technology
100 Bureau Drive, Stop 8910
Gaithersburg, MD, USA

Abstract – **A virtualization platform is responsible for allocation and aggregation of radio resources from different access technologies as well as the distribution of the total capacity among Virtual Network Operators (VNOs). The Radio Resource Management (RRM) employed by each VNO should comply with the requirements specified in the Service Level Agreements (SLAs) of each user. A joint admission control and resource management scheme based on proportionally fair rate allocation among different users was proposed in our previous publication. Although, all SLAs are satisfied in that scheme, users with vastly different QoS requirements might not necessary be treated fairly in terms of the allocated rates. This is especially the case when the available capacities of the VNOs cannot support the maximum requested rates for all such users. This paper attempts to overcome this weakness by replacing the proportional fairness strategy with a more general concept of utility-proportional fairness. The proposed approach is evaluated by simulations under increasing congestion scenarios and the results show improved fairness in the allocated rates.**

Keywords – Virtualization, distributed resource allocation, utility-proportional fairness, Service Level Agreements (SLA), admission control.

I. Introduction

S*ervice-oriented* architecture is expected to enable flexibility in sharing and utilization of network resources, wider range of customized services, along with a reduction in the capital and operational expenditures [1]. *Virtualization* supports service-oriented architecture through decoupling of the services and functionalities from the underlying Radio Access Networks (RANs) [2]. It enables the transformation of the physical infrastructure into multiple logical networks that can be shared among different Virtual Network Operators (VNOs). As such, VNOs do not need to own the infrastructure. Instead, they obtain the resources from a centralized virtualization platform and enforce their own service requirements and policies through the process of Radio Resource Management (RRM) [3].

The diversity in users' Quality of Service (QoS) requirements drives the emergence of resource slicing along with virtualization [4]. Performance optimization in virtualized Heterogeneous Networks (Het-Nets) not only optimizes the performance of various slices but also maximizes the utilization of the overall shared resources [5]. Scalability limitations of centralized RRM necessitate decentralized resource management [6], [7]. Authors in [8] have proposed a distributed RRM model for dense 5G networks based on non-cooperative game theory. While their approach achieves energy efficiency, it does not incorporate customized specifications and requirements of different services. An adaptive two-layer decentralized RRM with slow and fast timescales has also been presented in [9]; however, the methodology does not include network virtualization and slicing concepts. Authors in [10] propose another distributed RRM with a focus on multi-connectivity in 5G networks. Their approach aims at reducing the processing costs and signalling overhead; but does not consider the notion of RAN slicing, isolation, as well as service orientation.

In our previous publication, we proposed a joint admission control and RRM for virtualized RANs [1]. Here, we extend the proportionally fair rate allocation scheme in [11] to a more general utility-proportional rate allocation [12]. The intention is to address some of the observed shortcomings of proportionally fair allocation such as giving advantage to users with low bandwidth requirements [12]. Similar to [1], our proposed scheme maximizes the aggregate system utility using a two-stage distributed optimization on a *fast* and *slow* time scale and overcomes the scalability issues of the centralized RRM [13], [14]. At the faster time scale, and given the capacities of each VNO, users adjust their rates based on the congestion pricing. At the slower time scale, each VNO adjusts its own capacity according to its assigned congestion price subject to the total aggregate capacity of the system. The admission strategy, which requires limited degree of centralization, ensures system ability to guarantee minimum bandwidth requirements to the newly admitted user as well as to all users already present in the system.

The rest of this paper is organized as follows. Section II describes the system architecture and quantifies user preferences. Section III formulates system performance model. Section IV outlines the resource management scheme. Section V describes simulation scenarios and results. Finally, conclusion and plans for future research are discussed in section VI.

II. System Model

System architecture and quantification of user preferences by their corresponding utilities are described in the following subsections.

Marbukh, Vladimir; Sayrafian, Kamran; Rouzbehani, Behnam. "Utility Proportional Resource Allocation for Users with Diverse SLAs in Virtualized Radio Access Networks." Paper presented at IEEE International Black Sea Conference on Communications and Networking, Odesa, MD, US. May 26, 2020 - May 29, 2020.

A. System Architecture

Figure 1 shows the mechanism of service-oriented RAN slicing and resource management along with interaction of different entities in the system. The Virtual-RRM (VRRM) module is a centralized virtualization platform which is responsible for configuring the RAN protocol stack and QoS metrics according to the slice requirements. Those requirements are enforced by different VNOs based on their specific policies. As an example, assume that VNOs A and B provide two types of services with different requirements. For *slice* A with high throughput requirements, radio flow A is configured to support multi-connectivity. Therefore, slice A is using the resources from 2 different radio access points. On the other hand, the network *slice B* is configured with only one connection according to the provided policy.

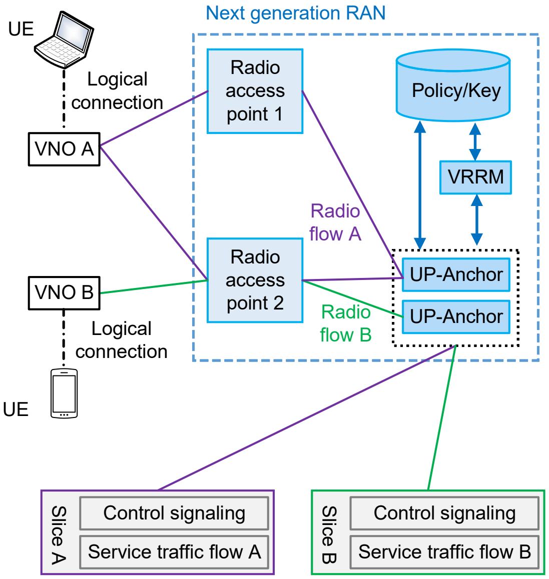


Figure 1. Service-oriented RAN slicing

The User Plane Anchor (UP-Anchor) is responsible for distributing the traffic flow in each slice. A RAN slice is composed of a control plane and a separate data plane. The required capacity allocation is subject to the SLA agreements between the VNOs and users. In this paper, we consider the following three categories of SLA contracts:

- Guaranteed Bitrate (GB): This is the highest priority category where a minimum threshold for data rate assignment must always be guaranteed regardless of the traffic load variation and network status. In addition, the assigned data rate need not exceed a maximum threshold for this SLA category.
- Best effort with minimum Guaranteed (BG): This is the second highest priority category for which a minimum level of data rate is guaranteed. Higher data rates are served in a best effort manner if available.
- Best Effort (BE): This is the lowest priority category for which there is no level of service guarantees and users are served in a pure best effort manner.

B. User Preferences

We assume that preference of each user for rate R can be quantified by the utility function $U(R)$, $R>0$. In [1], we assumed the following logarithmic utility:

$$U(R)=\lambda\log(\alpha R)\,, \tag{1}$$

However, logarithmic utility is typically inadequate for users with diverse QoS requirements which is the main driver of emerging resource slicing technology [12]. This utility function basically favors users with low bandwidth requirements, as observed in [12]. Logarithmic utility may also lead to negative utility values which could potentially cause undesirable oscillations during the rate allocation process. Some of these issues including negative utility values can be avoided by replacing utility (1) with the following utility function, shown in Fig.2.

$$U(R)=\lambda\log(1+\alpha R)\,. \tag{2}$$

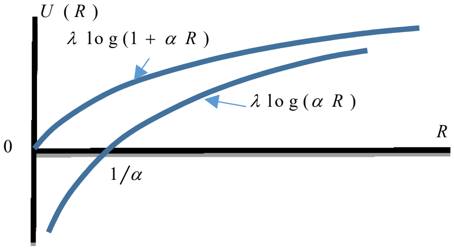


Figure 2. Logarithmic utilities

Fig. 3 exhibits the general utility of a user which requires certain minimum rate $R^{\min}$ and does not significantly benefit from rates above $R^{\max}$.

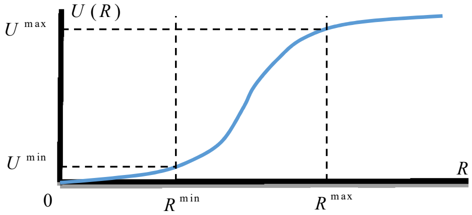


Figure 3. Utility of user with min/max rate guarantee

For example, GB users can be described by sigmoid utility $u(R)$ which is near zero for $R<R^{\min}$, quickly increases for $R^{\min}<R<R^{\max}$, and levels off for $R>R^{\max}$. A natural SLA for user with this utility is $R^{\min}<R<R^{\max}$. Sigmoid utility is often represented by the following function:

$$U(R)=\frac{k}{1+\exp[-\alpha(R-r)]}\,, \tag{3}$$

where parameters $k,\alpha,r>0$ can be expressed in terms of $R^{\min}$, $R^{\max}$, and $U^{\max}=U(R^{\max})$ [15]. Similarly, BG users are described by an utility which is near zero for $R<R^{\min}$, quickly increases for $R^{\min}<R<R^{\max}$, and then logarithmically increases for $R>R^{\max}$.

III. System Performance Optimization

The concept of utility proportional fairness and the class of utility functions which describes the SLAs considered in this paper are briefly discussed in the following subsections.

A. Utility Proportional Fairness

Let I_{sv} be the set of users obtaining service from the slice $s = 1,..,S$ of VNO $v = 1,..,V$, where sets I_{sv} with different (s,v) do not overlap, e.g., $I_{km} \cap I_{ln} = \emptyset$ if $(k,m) \neq (l,n)$, $k,l \in \{1,..,S\}$ and $m,n \in \{1,..,V\}$. Following Network Utility Maximization (NUM) framework [12], we assume that the goal of system management is maximization of the aggregate utility

$$U_\Sigma(R_i) = \sum\nolimits_{s=1}^{S} \sum\nolimits_{v=1}^{V} \sum\nolimits_{i \in I_{sv}} U_i(R_i) \tag{4}$$

over vector of rates (R_i) allocated to users $i \in I_{sv}$, $s = 1,..,S$; $v = 1,..,V$. This maximization is a subject to the following capacity and contractual constraints. The total capacity allocated to all users serviced by VNO v, $i \in I_{sv}$, $s = 1,..,S$ cannot exceed the VNO v capacity C_v:

$$\textstyle\sum_{s=1}^{S} \sum_{i \in I_{sv}} R_i \leq C_v, v = 1,..,V. \tag{5}$$

Also, the aggregate capacity allocated to all VNOs cannot exceed the total system capacity C^{VNNO}:

$$\textstyle\sum_{v=1}^{V} C_v \leq C^{VNNO}. \tag{6}$$

The above constraints are due to data rate guarantees to a user $i \in I_{sv}$ in slice s, i.e. R_s^{min} and R_s^{max} respectively:

$$0 \leq R_s^{min} \leq R_{svi} \leq R_s^{max},\ \ s = 1,..,S,\ \ v = 1,..,V \tag{7}$$

The second set of constraints is due to guarantees on the minimum capacity of each VNO v, $C_v^{min} \geq 0$:

$$C_v \geq C_v^{min}, v = 1,..,V. \tag{8}$$

Here, we consider a distributed solution to the aggregate utility (2) maximization:

$$\max_{(C_v)} \max_{(R_i)} \sum\nolimits_{s=1}^{S} \sum\nolimits_{v=1}^{V} \sum\nolimits_{i \in I_{sv}} U_i(R_i) \tag{9}$$

subject to constraints (5)-(8). Note that due to lower bounds in (7) and (8), optimization problem (5)-(9) may not have a feasible solution. This possibility necessitates an admission control similar to the process described in [1].

For concave user utilities, including logarithmic utilities (1) and (2), optimization problem (5)-(9) is convex; and therefore, the local maximum is also a global maximum (R_i^*). This is assuming that feasible sets (5)-(8) are non-empty. In the particular case of logarithmic utility (1), solution (R_i^*) is proportionally fair for any feasible allocation (R_i), i.e.

$$\sum\nolimits_{i \in I_{sv}} \lambda_i (R_i - R_i^*) / R_i^* \leq 0. \tag{10}$$

For non-concave user utilities (e.g., sigmoid utility (3)), optimization problem (5)-(9) is non-convex; and therefore, not generally tractable.

To resolve problems with proportional fairness, utility proportional fairness has been proposed in [12]. Rate allocation (R_i^*) is utility $u_i(R_i)$ proportional if

$$\sum\nolimits_{i \in I_{sv}} [(R_i - R_i^*) / u_i(R_i^*)] \leq 0 \tag{11}$$

for any feasible allocation (R_i). Proportional fairness (10) is a particular case of utility $u_i(R) = \lambda_i^{-1} R$ proportional fairness. It is known that utility $u_i(R)$ proportional fairness is equivalent to NUM with modified utility [12]

$$U_i(R) = \int_{R_i^{\min}}^{R} dr / u_i(r),\ R_i^{\min} \leq R \leq R_i^{\max}, \tag{12}$$

i.e., equivalent to aggregate utility maximization

$$\max_{(C_v)} \max_{(R_i)} \sum\nolimits_{s=1}^{S} \sum\nolimits_{v=1}^{V} \sum\nolimits_{i \in I_{sv}} U_i(R_i). \tag{13}$$

subject to constraints (5)-(8).

Note that the NUM problem should be solved every time the set of users changes due to user arrivals/departures. Assuming that resource optimization occurs on a faster time scale changes in the number of users, distributed solution to NUM (5)-(8), (12)-(13) is discussed in the next section. As an example, consider the following utility function,

$$u(R) = [u^{\min} + \lambda^{-1}(R - R^{\min})]_{u^{\min}}^{u^{\max}}, \tag{14}$$

where $\lambda = (R^{\max} - R^{\min}) / (u^{\max} - u^{\min})$, and $[z]_a^b = \max\{a, \min\{z,b\}\}$. Using utility (14) in equation (12) will lead to the following $U(R)$, and the comparison is shown in Figure 4.

$$U(R) = \left[\lambda \log\left(1 + \left(\frac{u^{\max}}{u^{\min}} - 1 \right) \frac{R - R^{\min}}{R^{\max} - R^{\min}} \right) \right]_0^{U^{\max}}, \tag{15}$$

where $U^{\max} = \lambda \log(u^{\max} / u^{\min})$.

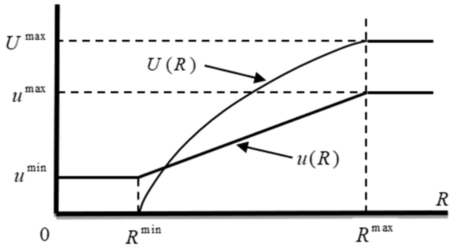


Figure 4. Piece-wise linear utility fairness

Our selection of utility (14) is due to its ability to describe BE users as well as users with rate guarantees.

IV. Resource Management

User rate and VNO capacity adaptation algorithms, given that the optimization problem (5)-(8), (12)-(13) has a feasible solution for the set of users, i.e., system has sufficient capacity to satisfy minimum rate requirements for all users in the system is presented in the following. The admission control strategy which basically ensures compliance with SLA for newly accepted as well as remaining users in the system is identical to the process described in [1].

User $i \in I_{sv}$ requests data rate by solving its individual optimization problem:

$$R_i(p_v) = \arg \max_{R_i^{\min} \le R \le R_i^{\max}} [U_i(R) - p_v R], \qquad (16)$$

where p_v is the price of a unit of data rate offered by the VNO v. Since function $U_i(R)$ is increasing and strictly concave for $R_i^{\min} \le R \le R_i^{\max}$,

$$R_i(p_v) = (1/k_i)\left([1/p_v]_{u_i^{\min}}^{u_i^{\max}} + k_i R_i^{\min} - u_i^{\min}\right). \qquad (17)$$

Figure 5 shows rate (17) versus price.

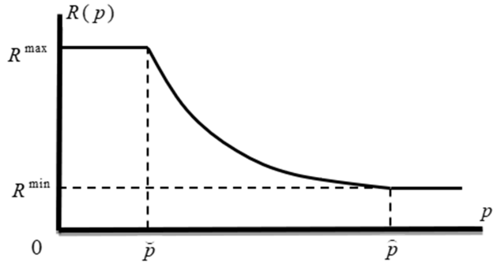


Figure 5. User rate vs. price

Due to the lower bound constraints in (7)-(8) optimization problem (5)-(8), (12)-(13) may not have a feasible solution. In this case, VNO v capacity deficit

$$\textstyle\sum_{s=1}^{S} \sum_{i \in I_{sv}} R_i^{min} - C_v > 0 \ \ , v = 1, .., V \qquad (18)$$

is arbitrarily allocated to currently present users in this VNO.

The optimal prices p_v^{opt} that maximize the utilization of the VNOs' available bandwidth are determined by the following distributed adaptive algorithm. The algorithm proceeds in discrete steps $k = \{1,2, \ldots\}$. At each step k, users solve the individual optimization problems (16) resulting in rate (17). If constraints (7)-(8) are satisfied, i.e., the aggregate data rate of the users does not exceed the total capacity of the associated VNO, then in step $k+1$ the price $p_{v,k+1}$ is reduced in order to motivate users to request higher rate. However, if the constraints (7)-(8) are not satisfied, $p_{v,k+1}$ is increased, resulting in a decrease of users' data rates. The main idea here is to maximize utilization of the available capacity in an efficient way. The price adaptation model can be expressed as [16]:

$$p_{v,k+1} = \left[p_{v,k} + h(\tilde{R}_{vk} - C_{vk})\right]^{+} \qquad (19)$$

where

$$\tilde{R}_{vk} = max\left(C_v^{min}, \textstyle\sum_{i \in I_{sv}} R_{ik}\right) \qquad (20)$$

$[x]^{+} = \max(0, x)$, and $h > 0$ is a small positive constant which regulates the tradeoff between optimality under stationary scenario and adaptability under non-stationary scenario, e.g., due to changing set of users. The main advantage of this approach is that VNOs do not have to know users' utilities which are considered private information.

In a slower time-scale each VNO adjusts its own capacity by negotiating the price with the VRRM. The adaptation of capacities among the tenant VNOs (C_v) is subject to the total available capacity of VRRM is C^{VRRM} (6). The average price of a unit of data rate in the entire system at step $k = \{1,2, \ldots\}$ is as follows:

$$P_k^{ave} = \frac{1}{C^{VRRM}} \sum_{v=1}^{V} C_v P_{v,k} \qquad (21)$$

where $P_{v,k}$ is the price of a unit of rate assigned to VNO v from VRRM at step k.

We propose the following capacity adaptation algorithm for the VNOs according to [16]:

$$C_{v,k+1} = C_{v,k} + H(P_{v,k} - P_k^{ave}), \qquad (22)$$

where $H > 0$ is a small constant.

Algorithm (21)-(22) increases (decreases) the capacity of a VNO if its corresponding price is higher (lower) than the average price (21). However, VNO capacity cannot fall below the lower bound in (8) due to equation (20).

V. Simulation Scenario & Results

To evaluate our proposed resource management strategy, the simple traffic distribution scenario with VRRM capacity of 510 Mbps has been considered in this section. Network parameters are defined in Table 1. It is assumed that 3 VNOs with different SLA types (i.e., GB, BG and BE) are providing services from 4 service classes: *Conversational* (Con), *Streaming* (Str), *Interactive* (Int.) and *Background* (Bac.) according to the class-of-service definition in UMTS. VNO GB delivers Voice (Voi), Video calling (Vic), Video streaming (Vis) and Music streaming (Mus). VNO BG serves File sharing (Fil), Web browsing (Web) and Social Networking (Soc) services, while VNO BE provides Internet of Things (IoT) and Email (Ema). It is further assumed that at each time step k, forty new users arrive and submit their requests for service to their associated VNOs. Simultaneously, twenty users depart from the system. For simplicity, the traffic type percentages of both arrivals and departures, defined as $U_{[\%]}^{srv}$, remain the same.

Table 1 – Network Parameters

VNO	**Service**	**Class**	$\boldsymbol{R_{svi}}$ in Mbps	$U_{[\%]}^{srv}$	$\boldsymbol{\lambda_s}$	C_v^{min} in Mbps
1 (GB)	Voi	Con.	[0.032, 0.064]	10	5	0.4 C^{VRRM}
	Vic		[1, 4]	10	4	
	Vis	Str.	[2, 13]	25	3	
	Mus		[0.064, 0.32]	15	1	
2 (BG)	Fil	Int.	[1, C^{VRRM}]	15	4	0.3 C^{VRRM}
	Web		[0.2, C^{VRRM}]	5	3	
	Soc		[0.4, C^{VRRM}]	10	2	
3 (BE)	Ema	Bac.	[0, C^{VRRM}]	5	4	0
	IoT		[0, 0.1]	5	4	

To evaluate the performance of user rate and VNO capacity adaptations, we consider stress scenario with proportionally increasing numbers of users of different services specified in Table 1. We assume user utility (2) as a particular case of utility-proportional rate allocation scheme. Benefits of utility-proportional fairness as compared to the previously used proportionally-fair strategy is demonstrated through extensive simulations. Figure 6 shows evolution of the system aggregate utility for utility proportionally rate allocation algorithm. The results show convergence of the users' rate adaptation and also highlights our assumption on separation of time scales, i.e., rate allocation should occurs much faster than changes in the set of users. The performance for the case where this assumption does not apply requires further investigation.

Figures 7 shows the converged system aggregate utility for utility proportional and proportionally fair resource allocation schemes. The advantage of utility-proportional scheme is clearly noticeable since admission of new users is inconsistent with decreasing aggregate utility. Figures 8 and 9 display another drawback of proportionally-fair rate allocation which gives advantage to users with lower rate SLA rate requirements. As the number of users increase at VNO GB, Voi users that have the lowest rate SLA requirements maintain their maximum requested rate of 0.064 Mbps long after Vis users rate drops to their minimum requested 2 Mbps. In general, under the proportionally fair resource allocation, users with highest SLA rate requirements will encounter reduced assigned rates well before users with lower SLA rates. Under a "fair" rate allocation scheme, all users should experience rate reduction from their maximum assigned rates approximately around the same time. As observed, this situation is better achieved using the utility-proportional strategy through proper adjustment of the parameter α in (2). Further results confirming this advantage have been omitted due to brevity.

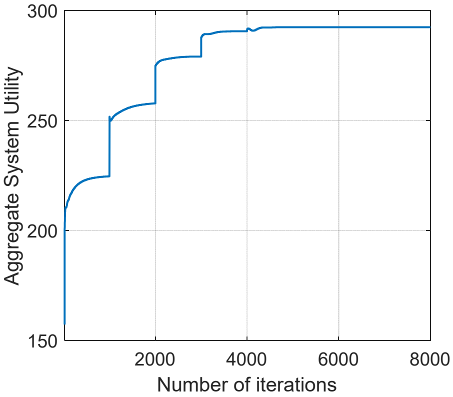


Figure 6. Evolution of aggregate utility for utility function (2)

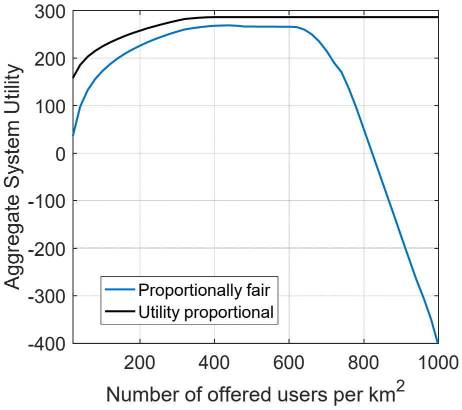


Figure 7. Aggregate utility for user utilities

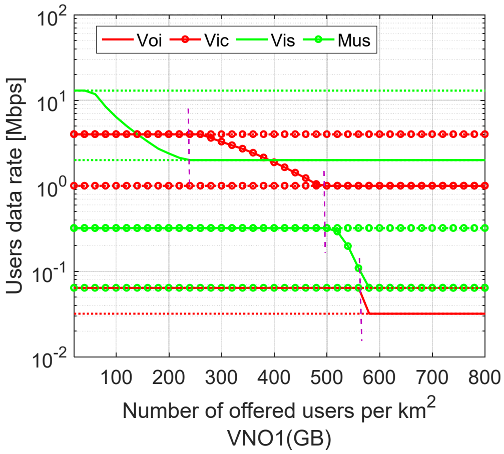


Figure 8. User rates for utilities (1)

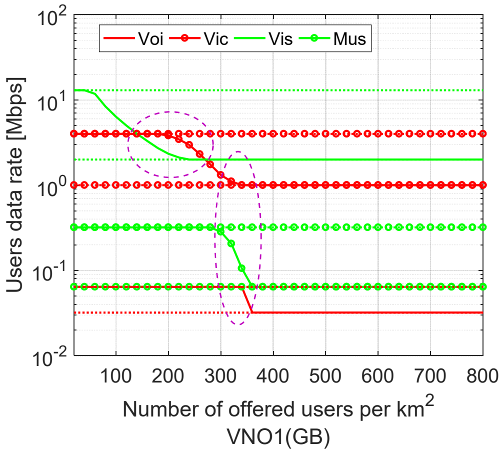


Figure 9. User rates for utilities (2)

VI. Conclusion and Future Research

The proposed radio resource management scheme in this research overcomes the shortcomings of the proportionally fair rate allocation by exploiting the general concept of utility proportional fairness. Simulation results clearly demonstrate advantages of using this strategy. In general, a customized SLA-based utility proportional fairness could lead to even better overall performance. The authors plan to further investigate this issue in future research. Viability of the time scale separation and convergence assumption in practical situations should also be studied. That will include mechanisms to mitigate performance loss in situations of comparable time scales in rate/capacity adaptation and users' arrivals/departures process. A requirement for this study is realistic models of users' arrival and departure processes. Finally, employing artificial intelligence (AI) techniques as a part of the network management could be a major focus of future research.

References

[1] B. Rouzbehani, V. Marbukh, and K. Sayrafian, "A Joint Admission Control & Resource Management Scheme for Virtualized Radio Access Networks", in *Proc. of CSCN'19 – 5th IEEE Conference on Standards for Communications and Networking*, Granada, Spain, Oct. 2019.

[2] Z. Feng, L. Ji, Q. Zhang and W. Li, “A Supply-Demand Approach for Traffic-Oriented Wireless Resource Virtualization with Testbed Analysis”, *IEEE Transactions on Wireless Communications,* Vol. 16, No. 9, Jun. 2017, pp. 6077–6090.

[3] M. Elkhodr, Q.F. Hassan and S. Shahrestani, *Networks of the Future: Architectures, Technologies, and Implementations*, CRC Press, Boca Raton, FL, USA, 2018.

[4] C. Liang and F. Yu, “Enabling 5G mobile wireless technologies”, *EURASIP Journal on Wireless Communications and Networking*, Vol. 2015, No. 218, Sep. 2015.

[5] A. Aijaz, “Towards 5G-enabled Tactile Internet: Radio Resource Allocation for Haptic Communications”, in *Proc. of WCNC’16 - 17th IEEE Wireless Communications and Networking Conference*, Doha, Qatar, Apr. 2016.

[6] S. Singh, S. Yeh, N. Himayat, S. Talwar, “Optimal Traffic Aggregation in Multi-RAT Heterogeneous Wireless Networks”, in *Proc. of ICC’16 –52th IEEE International Conference on Communications*, Kuala Lumpur, Malaysia, May 2016.

[7] M. Gerasimenko, D. Moltchanov and R. Florea, “Cooperative Radio Resource Management in Heterogeneous Cloud Radio Access Networks”, *IEEE Access*, Vol. 3, Apr. 2015, pp. 397–406.

[8] P. Sroka and A. Kliks, “Playing Radio Resource Management Games in Dense Wireless 5G Networks”, *Hindawi Journal of Mobile Information Systems*, Vol. 2016, Nov. 2016, pp. 1 – 18.

[9] F. Teng and D. Guo, “Resource Management in 5G: A Tale of Two Timescales”, in *Proc. of ACSSC’15 - 49th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA.

[10] V. Monteiro, D. Sousa and T. Maciel, “Distributed RRM for 5G Multi-RAT Multi-connectivity Networks”, *IEEE Systems Journal* (Early Access), Jun. 2018, pp. 1 – 13.

[11] B. Rouzbehani, V. Marbukh, K. Sayrafian, and L.M. Correia, “Towards Cross-Layer Optimization of Virtualized Radio Access Networks,” in *Proc. of EuCNC’19 – 28th European Conference on Networks and Communications,* Valencia, Spain, Jun. 2019.

[12] W.H. Wang, M. Palaniswami, and S. H. Low, “Application-oriented flow control: Fundamentals, algorithms and fairness,” *IEEE/ACM Transactions on Networking*, vol. 14, no. 6, pp. 1282 –1291, December 2006.

[13] B. Rouzbehani, L.M. Correia and L. Caeiro, “Radio Resource and Service Orchestration for Virtualised Multi-Tenant Mobile Het-Nets”, in *Proc. of WCNC’18 – 19th IEEE Wireless Communications and Networking Conference,* Barcelona, Spain, Apr. 2018.

[14] B. Rouzbehani, L.M. Correia and L. Caeiro, “A Fair Mechanism of Virtual Radio Resource Management in Multi-RAT Wireless Het-Nets”, in *Proc. of PIMRC’17 – 28th IEEE Symposium on Personal, Indoor and Mobile Radio Communications,* Montreal, QC, Canada, Oct. 2017.

[15] C. Liu, L. Shi, and B. Liu, Utility-Based Bandwidth Allocation for Triple-Play Services, Fourth European Conference on Universal Multiservice Networks (ECUMN'07).

[16] X. Lin, N.B. Shroff, and R. Srikant, “A Tutorial on Cross-Layer Optimization in Wireless Networks”, *IEEE Journal on Selected Areas in Communications*, Vol. 24, No. 8, August 2006.

MEASURED DATA ALIGNMENTS FOR MONITORING METAL ADDITIVE MANUFACTURING PROCESSES USING LASER POWDER BED FUSION METHODS

Shaw C. Feng, Yan Lu, and Albert T. Jones
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
Email: [shaw.feng, yan.lu, and jonesa]@nist.gov

ABSTRACT

The number and types of measurement devices used for monitoring and controlling laser powder bed fusion (LPBF) processes and inspecting the resulting AM metal parts have increased rapidly in recent years. The variety of the data collected by such devices has increased, and the veracity of the data has decreased simultaneously. Each measurement device generates data in a unique coordinate system and in a unique data type. Data alignment, however, is required before 1) monitoring and controlling LPBF processes, 2) predicting the material properties of the final part, and 3) qualifying the resulting AM parts can be done. Aligned means all data must be transformed into a single coordinate system. In this paper, we describe a new, general data-alignment procedure and an example based on LPBF processes. The specific data objects used in this example include in-situ photogrammetry, thermography, and ex-situ X-ray computed tomography (XCT), coordinate metrology, and computer-aided design (CAD) models. We use the data-alignment procedure to align the data from melt pool images, scan paths, layer images, XCT three-dimensional (3D) model, coordinate measurements, and the 3D CAD model.

Keywords: additive manufacturing, data alignment, data fusion, manufacturing system integration

1. INTRODUCTION

For three main measurement needs, accurately qualifying complex, laser powder bed fusion (LPBF)-built, metal parts is still extremely difficult. First, there are LPBF process instabilities, which can cause significant variations in built parts. Second, there are not well-understood quantitative relationships among the CAD geometry, the raw-material properties, the fabrication process, which are needed to predict final-part properties. Third, there are the current, LPBF-system states, which are needed for control. Meeting those three needs required the extensive use of a wide range of measurement devices, including sensors and measuring machines, to understand relationships before making decisions throughout the entire additively manufactured product life cycle.

High-quality parts require 1) understanding the physical phenomena, 2) developing the correct models of the different phenomenon, 3) linking those models through data, and, finally, 4) making the best life-cycle decisions possible. Meeting those four requirements were based on the measurement, the principle of physics, and associated mathematical models.

Metal additive manufacturing (MAM), is a completely new kind of fabrication technology. The life cycle of AM parts includes designing, engineering, controlling, and inspecting. Nevertheless, the required data links needed to perform those life-cycle functions optimally do not exist. Instead, LPBF MAM process users are building new kinds of data-driven models and information links, based on new kinds of data collected by new kinds of sensors. In addition to traditional numerical types of data, many AM sensors collect a myriad of different types and quantities of in-situ and ex-situ image data. For example, there are sensors for photogrammetry and thermography and machines for X-ray computed tomography (XCT) and coordinate measurement, such as coordinate measuring machines (CMMs) [9]. Unfortunately, different sensors collect data in different, physical, local, coordinate systems. Moreover, the data from sensors in those systems must be "fused" to create the links that provide input to the models needed to make all AM-part, life-cycle decisions.

Feng, Shaw C.; Lu, Yan; Jones, Albert T. "Measured Data Alignments for Monitoring Metal Additive Manufacturing Processes Using Laser Powder Bed Fusion Methods." Paper presented at the ASME 2020 International Design Engineering Technical Conferences & Computers & Information in Engineering Conference, St. Louis, MO, US. August 16, 2020 - August 19, 2020.

Today, aligning data is not possible because there are no procedures for both spatially and temporally aligning data from different coordinate systems. The scope of this work is on alignment of data. Data can come from scan commands, laser-spot locations, melt-pool images, layer-wise images, XCT model, CMM model, and CAD model. Our approach to aligning these different data types is to find and use the reference locations and orientations required to transform a local coordinate system into the reference coordinate system. The authors have identified the meta data needed to align related data sets, which can be used for data registration. Registered data can then be used for analyzing the quality of part.

This rest of paper has the following sections. Section 2 reviews related publications in data collected from both in-situ and ex-situ monitoring. Section 3 proposes a procedure for data alignment. Section 4 gives examples of data alignment. Section 5 discusses the proposed procedure. Section 6 concludes the paper and identify the future work.

2. REVIEW OF SENSOR DATA TYPES AND ALIGNMENT METHODS

Recently, researchers have been integrating LPBF sensors and developing techniques to apply sensor data as input to various applications, such as data analytics. The section provides a review of types of data, data fusion, and research needs.

2.1 Sensor Data Types

In-situ process monitoring is necessary for real-time control and is enabled by sensors that monitor in-process phenomena. Scime et al. [20] used a common, staring cameras with a k-Mean unsupervised classification algorithm to detect anomalies on freshly coated powder bed for laser PBF. The images are used to detect anomalies on a freshly coated powder bed surface: waviness from blade hopping, streaks, debris, voids, and incomplete spreading. Many researchers [11][12][3] have reported the use of multiple monitoring methods, including high-speed coaxial cameras, off-axis thermal detectors and/or staring cameras to collect data for monitoring melt-pool characteristics, including melt pool geometry, energy intensity, spatter, and residual heat.

Reutzel et al. [19] described measurements of melt-pool geometry and temperature with images taken by a single-color camera in the infrared (IR) range. Temperature measurements, however, were based on images taken from a dual-color camera. The authors aligned the images with built-in reference marks in addition to the part design. Everton et al. [5] described in-situ sensors and sensing techniques for monitoring part buildup, layer-by-layer, in LPBF processes. Purtonen et al. [17] applied optical sensors included photodiodes, spectrometers, Charged Coupled Device (CCD) and Complementary Metal Oxide Semiconductor (CMOS) imaging sensors, pyrometers, and infrared cameras to monitor the LPBF process.

Commonly used off-axis sensors are Digital Single Lens Reflex (DSLR) cameras. A DSLR camera can take images of the powder bed each time it is triggered. A combination of flashlights from different angles of illuminations can detect anomalies on each scanned layer [1]. Bartletta et al. [2] explored using a high-speed camera to record the laser-scanning process including melting, solidifying, and track formations. Foster et al. [6] used staring-video cameras and coaxial cameras to collect data for monitoring melt-pool characteristics with other types of sensors for data fusion to monitor the progress of melting and track formation.

2.2 Sensor Data Alignment

For the background, data alignment is part of the data registration process, as shown in Figure 1. Data registration is necessary to fuse different data correctly. Data alignment includes both temporal alignment and spatial alignment. Temporal alignment requires a synchronized clock, which usually is a prerequisite of spatial alignment. Spatial alignment is a process that converts sensor data from its original, local coordinate system to another coordinate system so that the data can be compared and fused with the data generated in the new coordinate system. AM data alignment as a research topic continues to expand. Nevertheless, the current limitations of that research are still impeding the use of advanced data analytics, which can accelerate the understanding and control of AM processes and improve decision-making across the AM part lifecycle. Specifically, data-correlation limitations include 1) a limited understanding of how to characterize the new types of sensor data and 2) there is no information to link the data correctly in space and time.

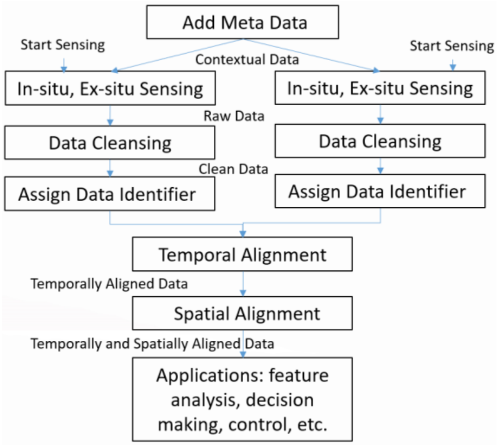


Figure 1 General Data Registration Procedure

2.3 State of Research in Sensor Data Alignment

Morgan et al. [13][14] used the direct-image-alignment approach, where no well-defined edges or corners in the build imagery can be used for alignment. This approach uses a filtered, built image and a synthetic image derived from the laser scan directions. These images are constructed so that they have high intensities where there is solidified materials and low intensities where there is unfused powder. The images must look as similar as possible so that they can be aligned by minimizing the differences. This requirement is almost impossible to be achieved due to measurement uncertainties associated with the sensor data.

Witherell [21] investigated data curation, fusion, and analytics techniques and showed an abstract data-alignment model for additive manufacturing. Bartlett et al. [2] correlated the measured surface temperature with layer-wise, scanned surface images to identify potential anomalies on the layer. The authors then correlated layer-wise measurements with the part measurements using ex-situ, scanning electron microscopy to validate the identified defects. Since the part used in the experiment is a regular shape, the data alignment is relatively simple.

Everton et al. [5] described specific sensors and sensing techniques for monitoring layer-by-layer builds in L-PBF processes. The purposes were to see 1) defects, such as pores, balling, unfused powder, and 2) cracks on a scanned surface with a correlation to the thermal images of that layer. Abdelrahman et al. [1] used a binary template created from the sliced 3D model of the part, as the reference geometry to do data alignment layer by layer. The aligned data then is used to create a 3D model to detect anomalies visually both before laser scanning and in the solidified material after laser scanning. Foster et al. [7] used angle illuminations to collect layer-wise images, which were then fused with pre-placed powder layers. Three-dimensional reconstruction of the images identifies potential flaws in the part.

Petrich et al. [16] used reference marks that were built into the part to align layer-by-layer images with an XCT model for defect location in the scanned part. Roehling et al. [18] modulated the heating and cooling profiles for visualizing the correlation between heating, cooling, and grain growth. Finally, Hirsch et al. [8] proposed a method to align 1) the design model, 2) the part slices, and 3) layer-by-layer images to create a 3D composite model for defect analysis.

2.4 Gaps and Research Needs

Clearly, there is an issue to properly relate a variety of sensors used for in-situ and ex-situ monitoring of LPBF AM processes. Those sensors provide a plethora of data, including gray-scale images and thermal data. While the data from individual sensors are important, correlations among those data can be extremely valuable. Sensor data must be aligned before the data can be applied for analysis to extract new knowledge. Data alignment is necessary to determine the state of the powder-fusion process, the material microstructure, and the fabricated part. For example, without correctly aligning measurements in the spatial domain, conflicting predictions can be made on the process performance and part quality. Lastly, there is no contextual information for data alignment. This is one of major barriers for part qualification and verification to ensure AM product quality

3. PROCEDURE FOR DATA ALIGNMENT

For in-situ monitoring, there are three sensor data types: photographic images, video clips, and time-series data. Photographic images include gray-scale images and thermographic images. Gray-scale images are commonly used to monitor melt pool shape, including its area and dimensions. Thermographic images are used to monitor temperature as well as energy intensity of a recently scanned layer. Video clips are used to monitor the dynamic behaviors in the scanning process, such as spattering and pluming. Time-series, acoustic data is collected from sonic sensors. Sonic sensor data is commonly used to monitor sparking or cracking during laser scanning.

For ex-situ monitoring, the following types of data are in the scope of this paper: XCT 3D model and points collected using CMM. CMM points should be properly associated with the corresponding features. An XCT 3D model can be used to identify pores and other defects. CMM points can be used to establish a datum reference frame and evaluate a feature's geometry again its tolerance specifications. These two types of data are commonly used in AM and are from nondestructive evaluations of additively manufactured parts.

Another type of data is scanning paths and speed, related to in-situ monitoring data. Scan paths are series of laser spot locations that are used to guide the laser to scan the powder layer. Lastly, chamber monitoring data, such as environmental temperature, gas pressure, and CO_2 levels. Note that chamber monitoring data is out of the scope of this work.

The above-mentioned data types are related, but in different coordinate systems. Examples of different coordinate systems include 1) the CAD-modeling coordinate system, 2) an in-situ sensor coordinate system, 3) a laser coordinate system, and 4) a staring camera coordinate system. Developing a procedure to geometrically align related data types and tie them all to a common coordinate system is the main purpose of our research work. The basis of our proposed, geometric, data-alignment procedure is coordinate transformation. The same point in the space is transferred from one coordinate system to another one. In this paper, the local coordinate system is referred as the "from" coordinate system, and the new coordinate system to which the point is transferred is referred as the "to" coordinate system. For example, a melt-pool image, which is collected by a coaxial camera, can be transformed from the camera coordinate system to the laser-scanning path coordinate system.

Scanning paths can be further related to the layer images taken by a staring camera. Layer images can be related to the 3D model generated by XCT. XCT 3D model can be related to the CMM model generated by the measured points (a point cloud). The CMM model can be related to the geometric model generated by a CAD system. In that sequence of relationships, melt pool images, scan paths, layer images, XCT model, CMM model, and CAD model can be sequentially related in data alignment.

In summary, the procedure includes the following steps:

(1) Group all the related data sets for data alignment. Specifically, two related data sets must have a common reference point and known difference in orientations, see Section 4 for examples.
(2) Identify every two related data sets and pair them for coordinate transformation. Also, identify the "from" coordinate systems and the "to" coordinate system in every pair.
(3) Sequence (chain) all the pairs according to spatial and/or temporal relations. For examples of spatial relations, see Section 4.
(4) Perform coordinate transformations so that all the data sets are in one single coordinate system.
(5) The chained data sets can be assigned an identifier (ID) for indexing and searching. The chained data sets can be used, such as data analysis.

4. EXAMPLES OF DATA ALIGNMENT

This section provides some data alignment examples. The sequence of align related data sets are as follows: (1) align melt-pool images to scan path, (2) align scan paths to the layer image, (3) align layer images to the XCT 3-D model, (4) align the XCT 3-D model to the CMM model, and (5) align the CMM model to the CAD model. After alignment, melt pool images, scanning paths, layer images, the XCT 3-D model, the CMM model, and the CAD model are all in the same coordinate system. This coordinate system will be the CAD coordinate system.

4.1 Melt pool image to scan path alignment

Melt pool is generated by laser melting. As shown in Figure 2, the center of the laser spot is used as the reference point in the alignment. The point in the coaxial camera coordinate system can be estimated using the shape of the melt pool shown in the image. The center of the laser spot on the scan path is in the scanning laser coordinate system. There are at least three ways to describe a scan path: (1) the command position in the XY2-100 or G-code file [10] (Note that the command position and the true laser spot center are different), (2) the intercepted encoder position of the two galvanometers [4], and (3) using an interpretation method to predict the true laser position based on the scanning speed, laser on/off timing, and camera-triggered times. When the time that the melt pool image is taken by the camera, the true laser spot moves away from the original position. If the off-distance is very small, then it is negligible.

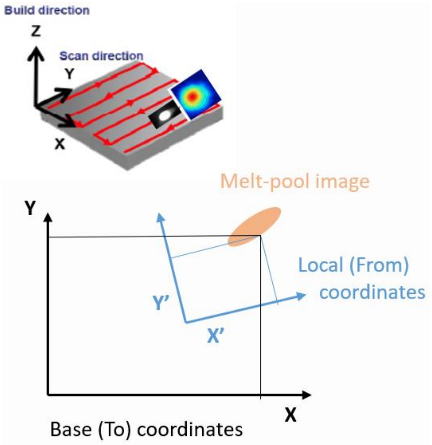


Figure 2 Melt pool images to scan path alignment

The relative orientation between the image ("from") and the layer ("To") can be computed using an appropriate image-calibration method. An image-calibration artifact with black and white grids may be used to measure the relative orientation difference between the orientation in the coaxial camera coordinate system and the orientation in the laser scanning coordinate system. The relations between the laser spot center and orientations in both coaxial camera and scanning laser coordinate systems are thus obtained for coordinate transformation. At this point, the melt-pool image is transformed to the scan path (layer) coordinate system.

4.2 Scan paths to layer image alignment

Scan paths should be aligned with the image of the layer that laser scanned. Scan paths are generated with the laser

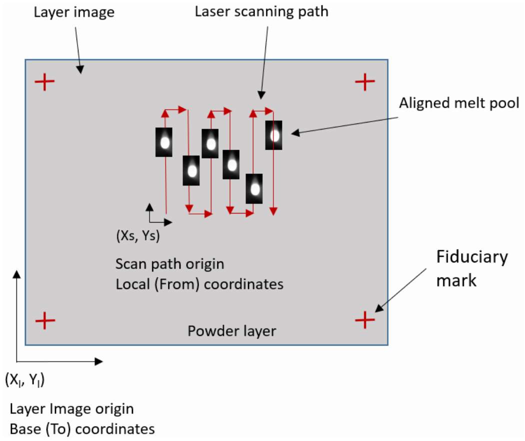


Figure 3 Scan path to layer image alignment

scanning coordinate system. Since it is not always possible to identify features on the scanned layer, fiduciary marks must be created. A fiduciary mark is created on the layer but outside the designed part during or after the laser scanning process. Four fiduciary marks can be used, minimally three. These four marks are the same artifacts in both scanning paths and the layer image. Figure 3 show an example of four fiduciary marks relative to the scan paths. With this aligning, coordinate transformation becomes possible. The "from" coordinate system is the laser scanning coordinate system. The "to" coordinate system is the layer image coordinate system. After the coordinate transformation, melt pool images, scan paths, and the layer image are all in the same coordinate system. Since there are multiple layers in an additively manufactured part, the fiduciary marks can also be used to align layers, from Layer 1 (the bottom layer) to the last Layer (the top layer), as shown in Figure 4.

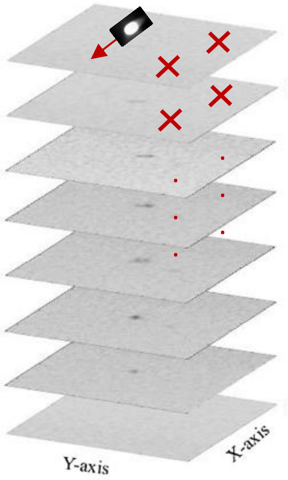


Figure 4 Layer images alignment

4.3 Layer images to the XCT model alignment

Aligned layer images should be aligned with the XCT 3-D model of the additively manufactured part. The purpose is to relate defects found in the XCT 3-D model to the defects found in the scanned layers to identify possible causes for those defects. There are some means for alignment: (1) create or specify reference datum features (e.g., plane, point, or line) in the part for alignment, (2) create fiduciary marks on last layer of the workpiece as the reference positions and align the fiduciary marks on the last layer image with the fiduciary marks shown in the XCT 3-D Model (note: the fiduciary marks has to be on the part, not outside the part so that XCT can detect them.), and (3) use mathematical algorithms to best fit layer images to the XCT 3-D model. The first method is based on geometric dimensioning and tolerancing standards, such as ANSI Y14.5. Figure 5 shows the fiduciary marks must be on the top of the part. The bottom few layers are not part of the part and are separated from the part when it is removed from the built plate.

4.4 XCT 3-D model to CMM model alignment

The XCT 3-D model should be aligned with the CMM model of the AM part. The purpose is to relate the part geometry found in the XCT model to the part geometry found in the CMM model to measure functional features, such as internal holes and thin walls, to verify if they are within the tolerances to ensure the manufactured part meets the functional requirements. The steps in the procedure includes (1) define a datum reference frame on the part for establishing a coordinate system, as defined in ANSI/ASME Y14.5, (2) align the XCT 3-D Model with the CMM model using the datum reference frame. Note that (1) a datum reference frame consists of primary, secondary, and tertiary reference planes (or equivalent geometries), (2) if datum

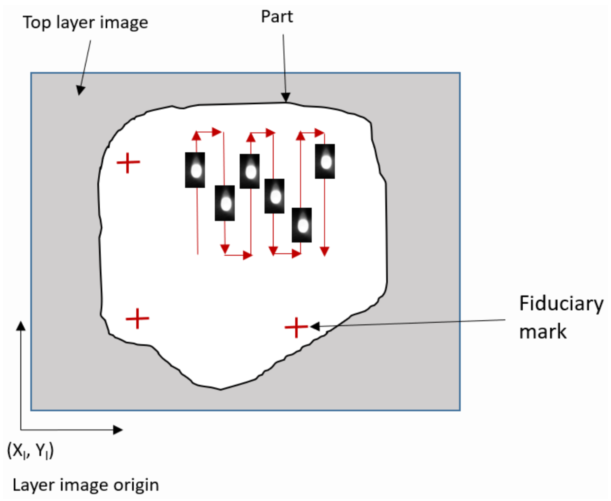


Figure 5 Layer image to XCT 3-D model alignment

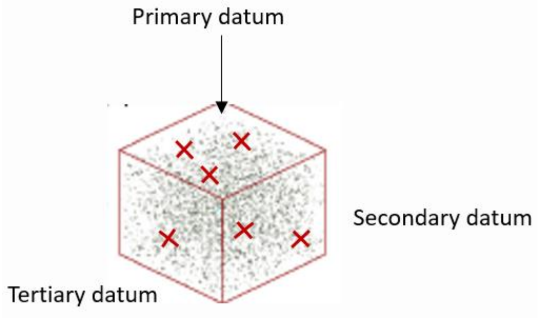


Figure 6 XCT 3-D model to CMM model alignment

reference frame is not possible to define, then other methods, such as fitting with point cloud, can be used for the alignment, and (3) CMM can be substituted with other dimensional measurement methods, such as laser scanners. Figure 6 shows an example of primary, secondary, and tertiary datums to form a datum reference frame, as specified in ANSI/ASME Y14.5.

4.5 The CMM model to the CAD model alignment

The CMM model should be aligned with the CAD model of the AM part. The purpose is to relate the part geometry found in the CMM model to the part geometry found in the CAD model to verify whether features with the specified tolerances, such as cylindricity of an internal hole, to verify if the features meet the tolerance requirements. The step in the procedure is to use the defined datum reference frame align the CMM model with the CAD model of the part for establishing a coordinate system, as defined in ANSI Y14.5. Figure 7 shows the use of primary, secondary, and tertiary datums to align CMM model to the CAD model. CMM model coordinate systems is the “from” coordinate system and the CAD coordinate system is the “to” coordinate system. With the alignment, the dimensions and tolerances of all the feature can be evaluated.

5. DISCUSSIONS

The primary challenge for AM is to control the fabrication process well enough to provide the reliability and repeatability necessary for commercial applications. For example, control is not a standalone process; control is connected to design/engineering processes upstream and inspection/qualification processes downstream. The key to the execution of all these AM processes is sensor data.

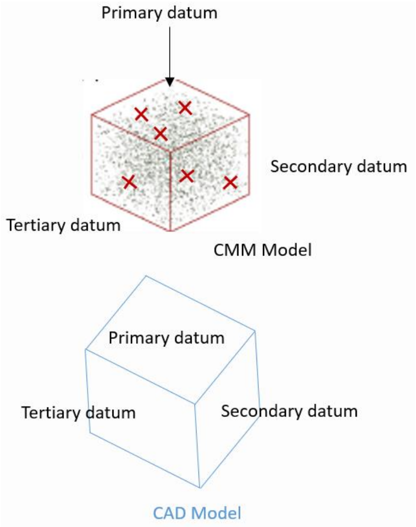


Figure 7 CMM model to CAD model alignment

In pursuit of overcoming the challenges, methods and standards of data gathering, registration, and fusion are the critical needs [15]. First, data should be collected and curated with rich meta information, e.g., sensor meta data, installation information, and configuration information. Additionally, best practices should be developed to calibrate the measurement apparatus and document the results appropriately. In addition, the data captured should be structured and represented to support interoperability among various computer information systems owned by various stakeholders, including material suppliers, machine manufacturers, measurement devices providers as well as testing labs. Both lexicon and semantic standards are required to enable seamless integrations of data generated from AM lifecycle and value chain activities. With established common data dictionary and common data exchange format, data collected during AM processes can be aligned and fused for better process monitoring and process control. In addition, data of thousands of builds conducted distributed can be aggregated into a common data virtual repository in the form of a federated data repository which will be available for the AM community to conduct advance data analytics including data alignment and data fusion and adaptive learning to accelerate AM part development lifecycle. This research establishes a good foundation on identification of the information for data fusion and provides a initial guidance on how to align the data. The remaining challenges include 1) data alignment for unsynchronized data, 2) data alignment uncertainty qualification, and 3) geometric feature alignment.

The data alignment procedure proposed in this paper is still a conceptual exploration. Real cases of additively manufacturing part using LPBF with in-situ and ex-situ measurements using appropriate sensors should take place to validate the procedure. Furthermore, defect-detection and cause analyses should also be done using the aligned data sets.

6. CONCLUSIONS

The use of laser powder bed fusion, LPBF, additive manufacturing technology to fabricate complex, metal parts in aerospace and medical industries has been increasing steadily. As a result, the demands on the quality and reliability of those parts has also increased. To respond to these demands, researchers have started to implement in-situ sensors and ex-situ measurement machines to monitor LPBF processes and to detect potential anomalies in the part. Types of in-situ data include melt-pool images, movies, and acoustic signals. Types of ex-situ data include XCT 3-D models and CMM data clouds. Correlation among related datasets are critical to detect anomalies and the causes. Correlated data set can also be used to verify the quality of AM parts. Since data alignment means all data is transformed into a single coordinate system, to compute the correlation is then possible for data analytics.

The proposed data alignment procedure in this paper addresses the long standard issue of how to align images taken by a coaxial camera, laser scanning commands, and a staring camera that are related to in-situ monitoring. The proposed procedure also addresses the need of aligning in-situ monitoring data with ex-situ monitoring data from XCT, coordinate measurement, and the design model. Examples in the paper show how to relate in-situ and ex-situ data into a suite of correlated datasets that can be used for downstream applications, such as defect analysis, feature analysis, and decision making.

Future work will be in two areas. One is to provide more examples of aligning time series data, such as acoustic signals, with geometric data. Second is to develop data registration procedure to include sensor meta-data with the sensor data for AM data analytics. The procedure will lead to standardization. We expect that these standards will lead to better implementations in the L-PBF user community. Furthermore, a case study that includes design, build, measurement, test should show that data alignment can enable defect detection in AM parts.

DISCLAIMER AND ACKNOWLEDGEMENT

Certain commercial equipment, instruments, or materials identified in this paper are not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

We also feel thankful for the support and the insightful discussions provided by Dr. Paul Witherell and Dr. Vijay Srinivasan at NIST.

REFERENCES

[1] Abdelrahman, M., Reutzel, E., Nassar, A., and Starr, T., “Flaw Detection in Powder Bed Fusion Using Optical Imaging,” Journal of Additive Manufacturing, Vol. 15, 2017, pp. 1 – 11.

[2] Bartletta, J., Heima, F., Murty, Y., and Lia,X., “In situ defect detection in selective laser melting via full-field infrared thermography,” Journal of Additive Manufacturing, Vol. 24, 2018, pp. 595 – 605.

[3] Chen, B., Lydon, J., Cooper, K., Cole, V., Northrop, P., and Chou, K., “Melt pool sensing and size analysis in laser powder-bed metal additive manufacturing,” Journal of Manufacturing Processes, Vol. 32, 2018, pp. 744- 753.

[4] Dunbar, A., Nassar, A., Reutzel, E., Blecher, J., “A Real-time Communication Architecture For Metal Powder Bed Fusion Additive Manufacturing,” Solid Freeform Fabrication 2016: Proceedings of the 27th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, 2016.

[5] Everton, S., Hirsch, M., Stravroulakis, P., Leach, R., and Clare, A., “Review of In-situ Process Monitoring and In-situ Metrology for Metal Additive Manufacturing,” Journal of Materials and Design, Vol. 95, 2016, pp. 431 – 445.

[6] Foster, B., Reutzel, E., Nassar, A., Dickman, C., and Hall, B., “A Brief Survey of Sensing For Metal-based Powder Bed Fusion Additive Manufacturing,” Dimensional Optical Metrology and Inspection for Practical Applications IV, edited by Harding, K., Yoshizawa, T., Zhang, S., Proceedings of SPIE, Vol. 9489, 2015, doi: 10.1117/12.2180654.

[7] Foster, B., Reutzel, E., Nassar, A., Hall, B., and, Dickman, C., “Optical, Layerwise Monitoring of Powder Bed Fusion,” Proceedings of the 26th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, 2015, pp. 295 – 307.

[8] Hirsch, M., Patel, R., Li, W., Guan, G., Leach, R., Sharples, S., and Clare, A., “Assessing the capability of in-situ nondestructive analysis during layer based additive manufacture,” Journal of Additive Manufacturing, Vol. 13, 2017, pp. 135 – 142.

[9] Hocken, R., et al., Coordinate Measuring Machines and Systems, CRC Press, 2016.

[10] Lane, B. and Yeung, H., “Process Monitoring Dataset from the Additive Manufacturing Metrology Testbed (AMMT): Three-Dimensional Scan Strategies,” Journal of Research of the National Institute of Standards and Technology, Vol. 124, Article No. 124003, 2019,

[11] Montazeri, M., et al., “In-Process Condition Monitoring in Laser Powder Bed Fusion,” SFF Symposium, 2017.

[12] Montazeri, M. et al., “Sensor-Based Build Condition Monitoring in Laser Powder Bed Fusion Additive Manufacturing Process Using a Spectral Graph Theoretic Approach,” Manuf Science and Engineering, 2018.

[13] Morgan, J., Morgan, P., Natale, P., Smith, R., Mitchell, W., Dunbar, A., and Reutzel, E., “Selection and Installation of High Resolution Imaging to Monitor the PBFAM Process, and Synchronization to Post-Build 3D Computed Tomography,” Proceedings of the 28th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, 2017, pp. 1382 – 1399.

[14] Morgan, J., “Data Fusion For Additive Manufacturing Process Inspection,” Master Thesis, The Pennsylvania State University, 2019.

[15] National Academies of Sciences, Engineering, and Medicine, “Data-Driven Modeling for Additive

Manufacturing of Metals: Proceedings of a Workshop," Washington, DC: The National Academies Press, 2019, https://doi.org/10.17226/25481.

[16] Petrich, J., Gobert, C., Phoha, S., Nassar, A, and Reutzel, E., "Machine Learning for Defect Detection for PBFAM Using High Resolution Layerwise Imaging Coupled With Post-Build CT Scans," Proceedings of the 28th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference, 2017, pp. 1363 – 1381.

[17] Purtonen, T., Kalliosaari, A., and Salminen, A., "Monitoring and Adaptive Control of Laser Processes," Physics Procedia, Vol. 56, 2014, pp. 1218 – 1231.

[18] Roehling, T., et al., "Modulating laser intensity profile ellipticity for microstructural control during metal additive manufacturing," Acta Materialia, 2017.

[19] Reutzel, E. and Nassar, A., "A survey of sensing and control systems for machine and process monitoring of directed-energy, metal-based additive manufacturing," Rapid Prototyping Journal, Vol. 21 Issue: 2, 2015, pp.159-167, https://doi.org/10.1108/RPJ-12-2014-0177.

[20] Scime, L. and Beuth, J., "Anomaly detection and classification in a laser powder bed additive manufacturing process using a trained computer vision algorithm," Journal of Additive Manufacturing, Vol. 19, 2018, pp. 114 – 126.

[21] Witherell, P., "Emerging Datasets and Analytics for Additive Manufacturing," Proceedings of Fraunhofer Direct Digital Manufacturing Conference DDMC 2018, Fraunhofer Verlag, Berlin, Germany, March 14 – 15, 2018, pp. 43 – 48.

Presented at the 95th ARFTG Microwave Measurements Conference, June 2020

Over-the-Air Testing of Cellular Large-Form-Factor Internet-of-Things Devices in Reverberation Chambers*

Kate A Remley[1], Clive Bax[2], Edwin Mendivil[3], Michael Foegelle[3], John Kvarnstrand[4], Derek Skousen[4], David A. Sánchez-Hernández[5], Miguel Garcia-Fernandez[5], Lorien Chang[6], Elvis Yen[7], Johnny Gutierrez[7], and Justin Harbour[7]

[1]National Institute of Standards and Technology, Boulder, CO USA; [2]Bureau Veritas; [3]ETS-Lindgren, [4]Bluetest AB, [5]EMITE, [6]Sporton, [7]Dell

***Abstract*—We demonstrate the applicability of the reverberation chamber for over-the-air testing of power-based metrics for cellular-enabled wireless internet-of-things devices. A recent industry-sponsored round robin illustrates agreement between labs to within the industry specified limits of ± 2 dB for radiated power and ± 2.3 dB for receiver sensitivity.**

***Index Terms* — cellular device testing; internet of things; over-the-air measurement; wireless system.**

I. Introduction

Over-the-air (OTA) testing is an important facet in the design and performance verification of wireless devices with integrated antennas [1]. We illustrate the applicability of the reverberation chamber for OTA testing with results from an international round robin focused on large-form-factor cellular-enabled internet-of-things wireless devices. Organized by the cellular industry association CTIA Certification™, the round robin concluded in January 2020 with measured data from six laboratories: United States (3), Europe (2), and Asia (1).

The round robin compared measurements performed in eight different reverberation chambers and three anechoic chambers of Total Radiated Power (TRP) and Total Isotropic Sensitivity (TIS). Tests of these "power-based" metrics are required for cellular industry certification. Measurements were carried out according to newly standardized procedures outlined in [2].

Two cellular devices were tested in four cellular bands with center frequencies ranging from 710 MHz to 2132 MHz and two radio-access technologies (Wideband-Code-Division Multiple Access W-CDMA, and Long-Term Evolution, LTE).

Results show agreement between measurements to within CTIA Certification's specified limits of ± 2 dB for TRP and ± 2.3 dB for TIS. Agreement was observed for all bands and radio-access technologies from physically diverse reverberation chambers. Agreement was also observed with the anechoic chambers. Such agreement between labs allows the CTIA Certification to move forward with authorization of OTA test laboratories who desire to utilize reverberation chambers.

II. OTA Testing of Cellular Wireless Devices

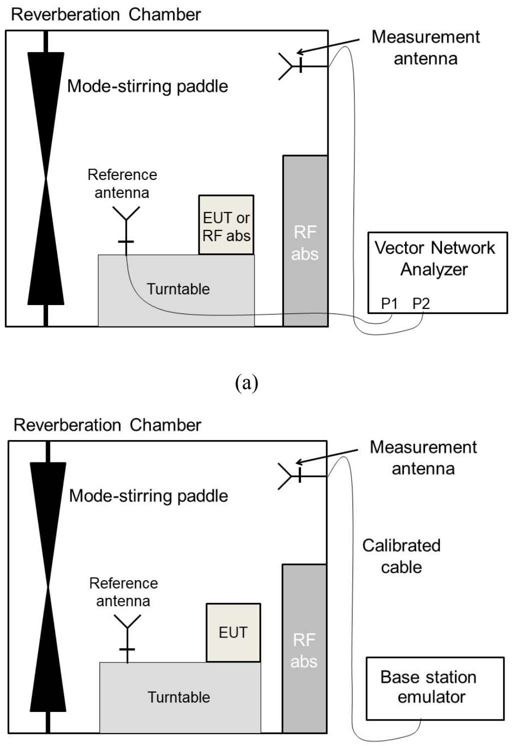
OTA testing in anechoic chambers has been standardized for many years [3]. In these tests, a three-dimensional characterization of the DUT's radiated performance is pieced together by analyzing data from spatially distributed measurements on a virtual sphere surrounding the device under test (DUT). Samples are spaced equidistantly along the theta (θ) and phi (ϕ) axes (15° for TRP, 30° for TIS) for each of two orthogonal polarizations [3].

To correct for test system effects such as chamber path loss, gain of the measurement antenna, cable losses, and so forth, a reference measurement is performed. A reference antenna is mounted at the center of the anechoic chamber as a substitute for the DUT. A vector-network-analyzer (VNA)-based reference measurement is performed for each possible signal path to the measurement equipment. When combined with the gain of the reference antenna, this measurement relates the measured quantity to that from an isotropic radiator.

During OTA testing, a base-station emulator (BSE) controls the DUT. Connected to a measurement antenna located within the chamber, the BSE, requests that the DUT transmit at full power. For LTE systems, resource-block allocations are specified. For the tests in this study, the transmitter and receiver operate at different frequencies within the selected band. In addition to control, the BSE may also measure the power received from the DUT (for TRP measurements) or the DUT's reported error rate for base-station power levels that are incrementally reduced until a target error rate is observed (for TIS). When averaged in post processing, the data provide an estimate of the metric of interest.

A reverberation chamber may be used instead of an anechoic chamber for isotropic OTA tests. The reverberation chamber is an electrically large, reflective cavity that supports many modes. The fields corresponding to these modes add constructively and destructively at the receive antenna. The received power can vary significantly as a function of frequency and location of the receive antenna within the chamber. To perform a measurement, the modes are intentionally randomized by altering the metallic boundary conditions within the chamber or by probing the fields at different locations and/or polarizations within the chamber. The former is often accomplished by use of a mechanical mode-stirrer such as a rotating paddle, while the latter by use of a rotating platform or the use of multiple measurement antennas having different physical locations and polarizations.

Remley, Kate; Bax, Clive; Mendivil, Edwin; Foegelle, Michael; Kvarnstrand, John; Skousen, Derek; Sanchez-Hernandez, David; Garcia-Fernandez, Miguel; Chang, Lorien; Yen, Elvis; Gutierrez, Johnny; Harbour, Justin. "Over-the-Air Testing of Cellular Large-Form-Factor Internet-of-Things Devices in Reverberation Chambers." Paper presented at 95th ARFTG Microwave Measurements Conference, Los Angeles, CA, US. June 26, 2020 - June 26, 2020.



(a)

(b)

Fig. 1: Typical measurement set-up in a reverberation chamber for (a) characterizing the chamber's power transfer function and (b) measuring a DUT's ("EUT") total radiated power and total isotropic sensitivity. RF absorber ("RF abs") has been included to broaden the chamber's coherence bandwidth (from [2], used with permission).

For OTA tests of cellular devices in reverberation chambers, the power (for TRP) or error rate (for TIS) is sampled under a number of static, ideally uncorrelated "mode-stirring states." Statistical analysis is performed to determine whether a sufficient number of mode-stirring states have been obtained to provide an estimate of the quantity of interest to within a desired level of uncertainty.

When it is necessary to demodulate a received signal in order to estimate the error rate, as for TIS, "loading" of the chamber with RF absorber is typically required to flatten the frequency response of the chamber. This creates an RF environment in which the equalizers of the DUT were designed to operate, allowing the device to be tested without the chamber environment itself introducing errors into the measurement. A metric to evaluate the flatness of the channel response is the coherence bandwidth (CBW) [4], which, in [2], is increased through loading with RF absorber until it exceeds the channel bandwidth of the signal being measured.

In [2], identification of the proper loading and its impact on uncertainty due to the reduced spatial uniformity of the average power measured within the chamber is carried out in a precharacterization step. This is discussed further in the next section. Other precharacterization steps in [2] ensure that the antenna placement is free from significant coupling to the RF absorber and that the frequency step used to assess the chamber set-up is adequate. As shown in Section III, the agreement between participating labs in the round robin indicates that the development and standardization of the precharacterization steps within the CTIA Certification Program Wireless-Internet-of-Things Subgroup's Reverberation Chamber Ad Hoc Group, are sufficiently rigorous and general to allow diverse reverberation chambers to be used in device assessment.

As with the anechoic-chamber method, a reference measurement is made to calibrate out system effects. The reference measurement provides an estimate of the "power transfer function" G_{ref} of the chamber set-up. A typical configuration for the measurement of G_{ref} is illustrated in Fig. 1(a). A reference antenna, whose radiation pattern characteristics are similar to those of the DUT [5], is used. The power transfer function is then estimated from VNA measurements averaged over a mode-stirring sequence as

$$G_{\text{ref},t} = \frac{\frac{1}{NF}\sum_{n=1}^{N}\sum_{f=1}^{F}|s_{21}(f,n)|^2}{\Gamma_{\text{meas}}\Gamma_{\text{ref}}\eta_{\text{meas}}\eta_{\text{ref}}}, \qquad (1)$$

where the averages are taken over N mode-stirring samples and F frequencies across the channel bandwidth to be tested. The variables Γ_x and η_x refer to the mismatch and radiation efficiency of antenna x, where x corresponds to the measurement or reference antenna. For this precharacterization step, measurements are averaged over T_{pre} = 12 "independent realizations" of the mode-stirring sequence, in which the same number of stirring-sequence samples is used (for example, 20 paddle steps and 10 turntable positions), but each measurement is spatially uncorrelated with those from the other 11 independent realizations. Correlation is checked by use of a cross-correlation method. The use of 12 spatially-uncorrelated calibration samples allows the user to estimate the uncertainty due to lack of spatial uniformity for a given chamber set-up and mode-stirring sequence. More details on this and the other precharacterization steps may be found in [2].

When DUT measurements are performed, a single reference measurement is made with the DUT in the chamber to account for a potential increase in the chamber loss due to the device. The DUT measurements are then conducted by replacing the VNA with a base station emulator, as shown in Fig. 1(b). The DUT is measured over the same mode-stirring sequence as was used in the precharacterization process, but typically for one independent realization only. The TRP is then computed as

$$P_{\text{TRP}} = \frac{1}{N}\frac{\sum_{n=1}^{N}P_{\text{meas}}(n)}{G_{\text{ref}}\Gamma_{\text{meas}}\eta_{\text{meas}}G_{\text{cable}}}, \qquad (2)$$

where G_{cable} refers to the VNA-measured loss (written here as a negative gain) associated with the cable between the base station emulator and the measurement antenna reference plane, as shown in Fig. 1(b). The TIS is found as

$$P_{\text{TIS}} = G_{\text{ref}}\Gamma_{\text{meas}}\eta_{\text{meas}}G_{\text{cable}}\left(\frac{1}{N}\sum_{n=1}^{N}\frac{1}{P_{\text{BSE}}(n)}\right)^{-1}, \qquad (3)$$

where the device's TIS is assessed at a base station emulator power P_{BSE} that induces a specified error rate, as defined in [3]. The round-robin tests discussed below incorporate the procedures outlined above.

III. The Round Robin Tests

A. Overview

Tests were conducted between June 2017 and May 2019. Two devices were tested by each lab: a smart phone supporting single-input, multiple-output (SIMO) operation, and a simulated large-form-factor IoT device consisting of the smart phone mounted on a lossy, large-form-factor polymer shipping case filled with RF absorber. Large, lossy DUTs are the most challenging for reverberation-chamber measurements as they increase the lack of spatial uniformity beyond that of the chamber loading. A representative set-up is shown in Fig. 2.

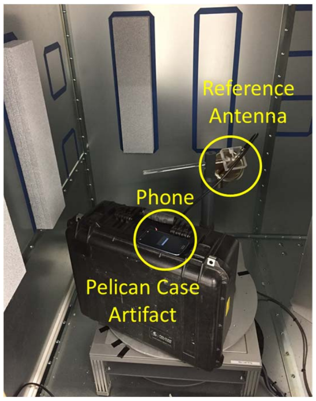


Fig. 2: The test artifact on a rotating turntable in a reverberation chamber. RF absorber (light gray) is magnetically mounted to the walls to broaden the coherence bandwidth. Display of product name does not imply endorsement by NIST. Other products may work as well or better.

The devices were measured in four licensed bands and two radio-access technologies, as specified in Table I. Per [3], TRP in Band 4 was measured over 12 resource blocks (RBs), while TIS was measured over both 50 and 100 RBs, with the latter under consideration for future versions of the CTIA test plan [3]. Labs conducted the entire set of measurements three days in a row. The smart-phone device exhibited unstable behavior in Band 17, which caused a delay until August 2017 when a second device was identified for use in that band.

Three of the eight reverberation chambers used in the round robin were deemed "small chambers," having precharacterized working volume dimensions less than 3 m on a side, while the other five were considered "large chambers." Testing was also carried out in three anechoic chambers although, as stated in Section II, a test methodology utilizing anechoic chambers for large-form-factor device tests is currently undefined in the CTIA Certification test plan. The lossy artifact with a single radiator in a known location presented little challenge for the anechoic chambers.

Table I: OTA tests conducted as part of the CTIA Certification round robin.

Band	Test	Mode	Channel	Frequency (MHz)	Bandwidth (MHz)
17	TRP	LTE	23790	710	2.4
17	TIS	LTE	5800	741	10
5	TRP	W-CDMA	4132	826.4	4
5	TIS	W-CDMA	4357	871.4	4
2	TRP	W-CDMA	9262	1852.4	4
2	TIS	W-CDMA	9662	1932.4	4
4	TRP	LTE	20175	1732.5	2.4
4	TIS	LTE	2175	2132.5	10
4	TIS	LTE	2175	2132.5	20

B. Precharacterization Results

Agreement between labs in this study is better than an initial round robin conducted by the same organization in the 2015-2016 timeframe. This is due to the extensive, newly finalized chamber precharacterization that is specified in the test plan [2]. The ability of each lab to meet the precharacterization specifications was checked.

As mentioned in Section II, the lack of spatial uniformity due to loading of the chamber can have a significant impact on the uncertainty in the measurement. In the test plan, this effect is quantified by taking the standard deviation of the T_{pre} = 12 reference power transfer function measurements from (1) as

$$\sigma_{G_{\mathrm{ref}}} = \sqrt{\frac{1}{T_{\mathrm{pre}}-1}\sum_{t=1}^{T_{\mathrm{pre}}}\left(G_{\mathrm{ref},t} - G_{\mathrm{ref}}\right)^2}. \quad (4)$$

Assuming a Gaussian distribution by the Central Limit Theorem, the standard uncertainty for the reference and DUT measurements is then calculated as

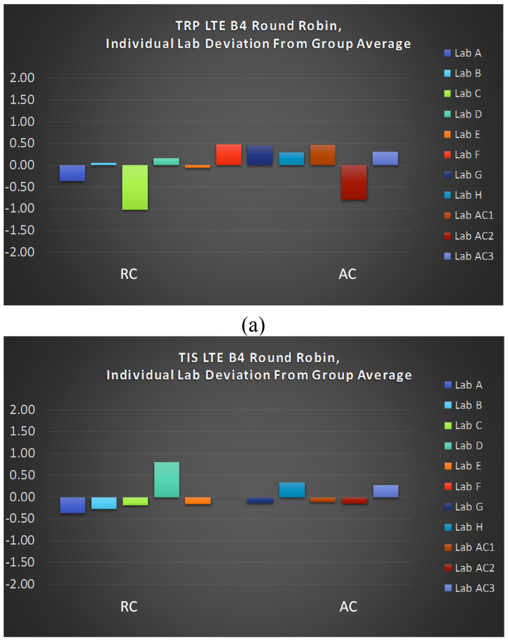
$$u_{G_{\mathrm{ref}}} = \frac{\sigma_{G_{\mathrm{ref}}}}{\sqrt{T_{\mathrm{cal}}}}, \quad (5)$$

where T_{cal} is the number of independent realizations of the reference power transfer function measured during the DUT measurement. Note that T_{cal} is often equal to one.

Laboratories generally need to keep the value of $\sigma_{G_{\mathrm{ref}}}$ as small as possible in order to meet the CTIA Certification combined uncertainty requirements of ± 2 dB for TRP and ± 2.3 dB for TIS. Minimizing this reverberation-chamber-specific component may generally be accomplished with the use of additional position and polarization stirring. While combined measurement uncertainty is not reported here, values of $\sigma_{G_{\mathrm{ref}}}$ ranged from 0.2 dB to 1.0 dB with typical values on the order of 0.5 dB to 0.6 dB.

C. DUT Results

CTIA Certification requires that the mean-normalized difference between labs falls below the uncertainty threshold in [3]. We first present results for the round-robin measurements of TRP and TIS from both the cellular smart phone alone in the chamber (Fig. 3) and the simulated large-form-factor IoT test artifact (Fig. 4). Results from Band 4 had the largest variation of any that was tested and, thus represent the worst case. In each



(a)

(b)

Fig. 3: Results for Band 4 LTE for the smart phone only: (a) TRP; (b) TIS. The highest standard deviation of any lab for three repeat measurements was 0.4 dB.

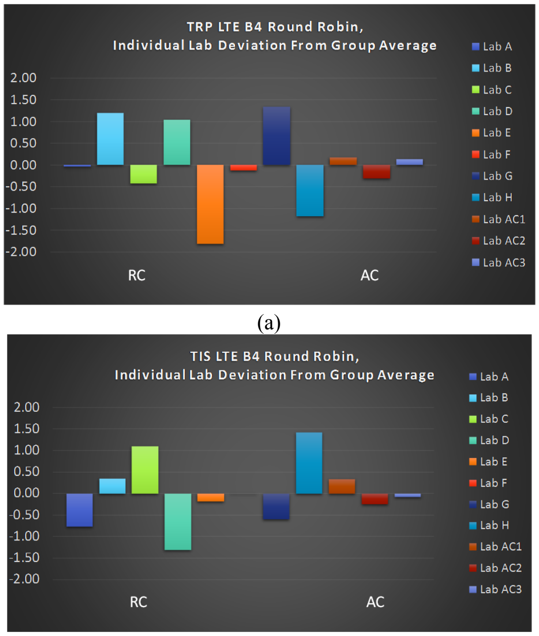
figure, the top graph shows TRP with f_c = 1.7313 GHz and a bandwidth of 2.4 MHz, while the bottom graphs show TIS with f_c = 2.1375 GHz and a bandwidth of 10 MHz. The three anechoic-chamber results are included at the right of each plot.

In each graph, the mean value (in decibels) of the respective chamber type (reverberation chamber or anechoic chamber) has been subtracted from the results. When the mean of both chamber types (anechoic-plus-reverberation chamber) is subtracted, results are typically within 0.2 dB of the reverberation-chamber-only mean, and always within 0.5 dB, indicating close agreement between chamber types.

For the smart-phone-only results in Fig. 3, no lab exceeds a 1.0 dB difference from another, and differences are typically less than 0.5 dB. This is because the DUT is small and does not significantly perturb the fields within the chamber. The results with the artifact in Fig. 4 exhibit a greater variation, with a maximum difference of 1.9 dB (the greatest difference in any band for the round robin), but differences are typically less than 1.0 dB. Some labs reported reduced repeatability for this DUT in Band 4, with Lab F unable to complete the TIS measurement possibly due to heating during prolonged testing.

IV. Conclusion

Round-robin agreement to within the CTIA Certification-specified limits between labs indicates that the newly standardized reverberation-chamber precharacterization and device measurement procedures are adequate to allow use of reverberation chambers for OTA testing of IoT devices incorporating the radio access technologies noted here.



(a)

(b)

Fig. 4: Results for Band 4 LTE with the large-form-factor artifact: (a) TRP; (b) TIS. The highest standard deviation of any lab for three repeat measurements was 0.7 dB.

Acknowledgement

Contributors to the test plan include NIST researchers Rob Horansky, Robert D. Jones, Stephan van de Beek, Willam Burger, Joop Aan Den Toorn, Jos Dortmans, Maria Becker and Matt North, Vincent Neylon and Eleazar Zuniga from Eurofins MET Labs. CTIA Certification's Sean Wang, Will Lightfoot, Marin Gavin and Mark Sargent also supported this work.

References

[1] Z. Liu, Y. Qi, F. Li, W. Yu, J. Fan, and J. Chen, "Fast band-sweep total isotropic sensitivity measurement," *IEEE Trans. Electromagnetic Compat.*, vol. 58, no. 4, Aug. 2016, pp. 1244-1251.

[2] CTIA Certification, "Test plan for wireless large-form-factor device over-the-air performance," v.1.2.1, Feb. 2019.

[3] CTIA Certification, "Test plan for wireless device over-the-air performance: method of measurement for radiated RF power and receiver performance," v.3.9, Nov. 2019.

[4] X. Chen, P.-S. Kildal, C. Orlenius, and J. Carlsson, "Channel sounding of loaded reverberation chamber for over-the-air testing ofwireless devices—Coherence bandwidth versus average mode bandwidth and delay spread," *IEEE Antennas Wireless Propag. Lett.*, vol. 8, pp. 678–681, 2009.

[5] K. A. Remley, R. J. Pirkl, C.-M. Wang, D. Senic, A. C. Homer, M. V. North, M. G. Becker, R. D. Horansky and C. L. Holloway "Estimating and correcting the device-under-test transfer function in loaded reverberation chambers for over-the-air tests," *IEEE Trans. Electromagnetic Compat.*, vol. 59, no. 6, Dec. 2017, pp. 1724 – 1734.

INVESTIGATING COUPLED EFFECT OF RADIATIVE HEAT FLUX AND FIREBRAND SHOWERS ON IGNITION OF FUEL BEDS – STRUCTURE vs VEGETATIVE FIREBRANDS-

Sayaka Suzuki[1] and Samuel L. Manzello[2]
[1]National Research Institute of Fire and Disaster, Japan
[2]National Institute of Standards and Technology, USA

1. INTRODUCTION

Large outdoor fires pose problems for societies across the world. Perhaps the most often in the news are wildland fires that approach urban areas. These are more simply referred to as Wildland-Urban Interface (WUI) fires. In Asia and North America, some recent examples are the 2019 WUI fires that occurred in South Korea and those in 2018 in Northern California in the United States. In African countries, there have been large fires that have occurred in informal settlements. For centuries there have also been large urban fires in Japan, a country with no large wildland fire problem or informal settlement situation [**1**].

In all of these large outdoor fires, firebrands are produced and lead to enhanced fire spread processes. Often referred to as spotting processes, firebrands are liberated from the combustion of various fuel types and then induce ignition of fuel sources away from the initial fire source. A potentially important aspect of the physics of ignition induced by firebrands is the coupled influence of firebrand showers and radiant heat from burning houses. To this end, a new experimental protocol was developed to study the coupled effect of radiation and firebrand showers on ignition processes of fuel beds [**2**]. Experiments were performed under an applied wind field, as the wind is a key parameter in large outdoor fire spread processes.

In this study, the type of firebrands was changed to study the effect of firebrand differences.

2. EXPERIMENTS

Details of experimental protocols are described in [**2**] hence only a short description is provided. To generate firebrand showers, the reduced-scale continuous-feed firebrand generator was used and installed inside NRIFD's wind facility. The device consisted of two parts: the main body and continuous feeding component.

A conveyer was used to feed wood chips continuously into the device. The conveyer belt was operated at 1.0 cm/s, and wood chips were put on the conveyer belt at 12.5 cm intervals. For all tests, Japanese Cypress wood chips were used to produce firebrands compared to Douglas-fir wood pieces used in [**2**]. These same size wood pieces have been shown to produce firebrands similar to the projected area/mass of firebrands produced by burning structures [**3**]. Here we call firebrands from Japanese Cypress wood chips structure firebrands, and firebrands from Douglas-fir wood pieces vegetative firebrands. The wood feed rate used here was 80 g/min, which is near the upper limit for this reduced-scale firebrand generator.

The blower was set to provide an average velocity below 4.0 m/s measured at the exit of the firebrand generator when no wood pieces were loaded. Above 4.0 m/s, smoke production was mitigated, but then many firebrands produced were in a state of flaming combustion as opposed to glowing combustion. In these experiments, glowing firebrands were desired.

The fuel beds used for ignition were 300 mm by 300 mm in size and consisted of Douglas-fir wood pieces with the dimensions of 7.9 mm x 7.9 mm x 12.5 mm. These were installed inside a mock-up corner assembly lined with calcium silicate board since the ignition of the corner assembly itself was not the goal here; only ignition induced in the wood pieces was of interest.

To provide uniform radiant heat flux to fuel beds, an electrically operated quartz radiant panel was used. Dimensions of the radiant panel were also 300 mm by 300 mm (**Fig. 1**). It was mounted at a height of 440 mm from the fuel bed surface. A custom calibration rig was designed and fabricated to quantify the radiant heat flux that the radiant panel provided at the fuel bed surface. The radiant heat flux at the fuel bed surface was 8.5 kW/m^2 with no wind. Pre-heating time was determined as the duration from the time when the radiant panel was turned on to the time to start the firebrand generator. Selected pre-heating time was 10 min. Baseline experiments were also performed to examine the ignitions by firebrands without applied radiant heat flux. The wind speed in this study was 6 m/s.

Manzello, Samuel; Suzuki, Sayaka. "Investigating Coupled Effect of Radiative Heat Flux and Firebrand Showers on Ignition of Fuel Beds - Structure vs Vegetative Firebrands." Paper presented at Japan Association for Fire Science and Engineering, Hirosaki, JP. May 28, 2020 - May 29, 2020.

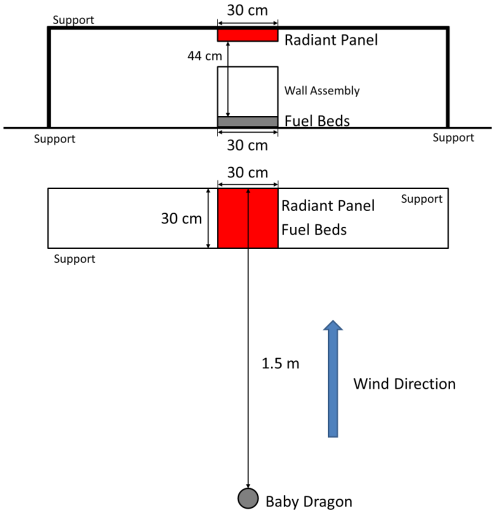


Figure 1 Schematic of the experimental settings. Front view (top) and Top view (bottom) are shown. Firebrand showers were produced using the reduced-scale firebrand generator installed in NIRFD's wind facility. Fuel beds consisted of Douglas-fir pieces of uniform size.

3. RESULTS AND DISCUSSION

Figure 2 shows the time to ignition (smoldering ignition, SI, and flaming ignition, FI) by structure firebrands compared with those by vegetative firebrands. For both type of firebrands, it required less time to ignite a fuel bed when radiate heat assistance was provided. Shorter times were needed to ignite a fuel bed by structure firebrands than by vegetative firebrands, regardless of external radiation. This is due to the fact that the contact area with a fuel bed is larger for structure firebrands than for vegetative firebrands. With 10 min pre-heating by radiative heat assistance, the time to ignition for both SI and FI reduced by 80 % for structure firebrands, while for vegetative firebrands, reduction was 65 % and 97 % for SI and FI respectively.

Figure 3 shows the mass required for ignition (unit area). The arrival firebrand mass flux of structure firebrands was 0.36 g/m^2 s and those of vegetative firebrand was 0.32 g/m^2 s, under 6 m/s wind condition. With firebrand mass flux considered, **Fig. 3** displays little difference from **Fig. 2**.

4. CONCLUSIONS

The coupled effect of radiative heat flux and firebrand showers on ignition processes of fuel beds was studied, focusing on the difference of firebrands. Structure firebrands ignited a fuel bed quicker than vegetative firebrands, regardless of radiative heat assistance. With 10 min pre-heating, the time to ignition for both SI and FI decreased dramatically for both structure and vegetative firebrands.

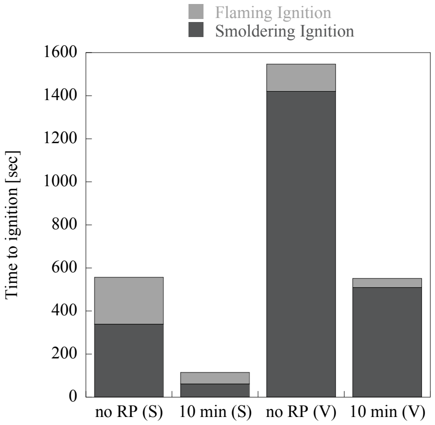


Figure 2 Time to ignition. No RP means baseline experiments where no radiant heat was provided. (S) and (V) means structure firebrands and vegetative firebrands respectively

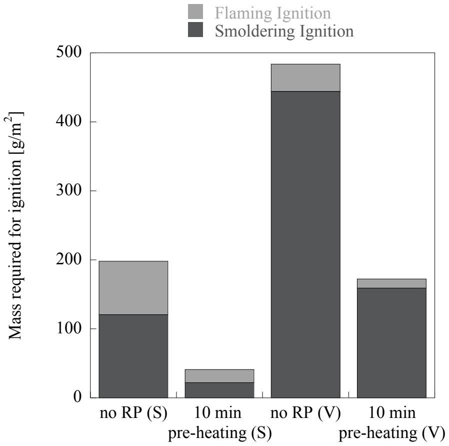


Figure 3 Mass required for ignition. No RP means baseline experiments where no radiant heat was provided. (S) and (V) means structure firebrands and vegetative firebrands respectively

5. ACKNOWLEDGEMENT

This research is supported by MEXT/JSPS KAKENHI Grant-in-Aid for Scientific Research (C) 18K04675.

REFERENCES

[1] Manzello, S.L., Suzuki, S., Gollner, M.J, and Fernandez-Pello, A.C., *Progress in Energy and Combustion Science*, (2020) 76, 100801. [2] Suzuki, S., and Manzello, S.L., *Proceedings of the 56th Japanese Combustion Symposium,* Tsukuba, Japan, 2019. [3] Manzello, S.L., Suzuki, S., *Proc. of the Combustion Institute,* (2017) 36(2) 3247-3252.

Manzello, Samuel; Suzuki, Sayaka. "Investigating Coupled Effect of Radiative Heat Flux and Firebrand Showers on Ignition of Fuel Beds - Structure vs Vegetative Firebrands." Paper presented at Japan Association for Fire Science and Engineering, Hirosaki, JP. May 28, 2020 - May 29, 2020.

HLVU : A New Challenge to Test Deep Understanding of Movies the Way Humans do

Keith Curtis
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
keith.curtis@nist.gov

George Awad*
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
george.awad@nist.gov

Shahzad Rajput*
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
shahzad.rajput@nist.gov

Ian Soboroff
National Institute of Standards and Technology
Gaithersburg, Maryland, USA
ian.soboroff@nist.gov

ABSTRACT

In this paper we propose a new evaluation challenge and direction in the area of High-level Video Understanding. The challenge we are proposing is designed to test automatic video analysis and understanding, and how accurately systems can comprehend a movie in terms of actors, entities, events and their relationship to each other. A pilot High-Level Video Understanding (HLVU) dataset of open source movies were collected for human assessors to build a knowledge graph representing each of them. A set of queries will be derived from the knowledge graph to test systems on retrieving relationships among actors, as well as reasoning and retrieving non-visual concepts. The objective is to benchmark if a computer system can "understand" non-explicit but obvious relationships the same way humans do when they watch the same movies. This is long-standing problem that is being addressed in the text domain and this project moves similar research to the video domain. Work of this nature is foundational to future video analytics and video understanding technologies. This work can be of interest to streaming services and broadcasters hoping to provide more intuitive ways for their customers to interact with and consume video content.

CCS CONCEPTS

• **Information systems → Retrieval tasks and goals**.

KEYWORDS

video understanding, multimedia, information retrieval, video ontology

ACM Reference Format:
Keith Curtis, George Awad, Shahzad Rajput, and Ian Soboroff. 2020. HLVU : A New Challenge to Test Deep Understanding of Movies the Way Humans do. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3372278.3390742

*Georgetown University

1 INTRODUCTION

Video understanding is a very difficult problem to solve. All current video analysis technology relies on detection, recognition and analysis of certain specific visual concepts such as people, objects, actions, activities or events. All these visual concepts recognition are usually done in isolation from the context of the video and only focusing on visual clues. This is one of the reasons why the current state of the art is lacking high level video understanding capabilities that connect the entities of a video with different events or relations. For example, given a two hour movie the current computer vision systems are not able to understand the relationship between different characters and develop a deep understanding of the video context. There has been efforts to encourage research in high level video understanding such as the "MovieQA" and "The Large Scale Movie Description Challenge" [9]. However these tasks revolve around isolated visual concepts retrieval and not about testing systems for their overall understanding of entities, relations and events within the video/movie. Early visions of this kind of work [7] proposed to use visual and audio descriptors, in addition to employing semantic analysis and linking with external knowledge sources in order to populate a knowledge graph.

The goal of our proposed research is to design and build datasets and evaluation benchmarks to foster the interest of the research community to develop systems which can extract available information from a video (e.g. a movie characters, their story lines, and relationships), and to use this information to reason about other, more hidden background information, and eventually to populate a knowledge graph with all extracted and reasoned information. In the next section we present an overview to related work in this area followed by more technical details discussing a pilot dataset collection of open source movies, the human annotation framework, query design and preprocessing, and our proposed evaluation roadmap and future directions. This is an ambitious new area of research which is to be run initially as a Grand Challenge at ACM Multimedia 2020 and as a workshop at ACM ICMI 2020 respectively. TRECVID workshop participants will be invited to participate in this new task.

2 RELATED WORK

Integrating vision and language research has recently gained a lot of attention to promote image and video understanding. Most

Curtis, Keith; Awad, George; Rajput, Shahzad; Soboroff, Ian. "HLVU: A new challenge to test deep understanding of movies the way humans do." Paper presented at ACM International Conference on Multimedia Retrieval (ICMR), Dublin, IE. June 08, 2020 - June 11, 2020.

approaches adopt the question answering paradigm as their evaluation framework. For example in [2] the authors propose the task of visual question answering for images. They provide a dataset containing 0.25M images, 0.76M question and 10M answers. However, most question types target very specific visual facets in the image such as what, where, number of objects, and what is the attributes or relation of an object to others.

MovieQA [11] is a dataset which aims to evaluate automatic story comprehension from video and text. It consists of 14,944 multiple choice questions, each with 5 multiple-choice answers, only one of which is correct, from about 408 movies with high semantic diversity. The movies have been segmented into video clips of maximum 200s durations where participants have to answer a question related to this video clip. The dataset itself comes with multiple answering sources for the questions such as plot synopses, scripts, subtitles, and audio descriptions. Annotators essentially used the plot synopses to come up with the set of questions and answers instead of watching the whole movie.

Following from this, [8] explored the biases in the MovieQA dataset and found that by using an appropriately trained word embedding, about half of the Question-Answers can be answered by looking at the questions and answers alone, completely ignoring the narrative context from video clips, subtitles, and movie scripts.

A large-scale dataset of corresponding movie trailers, plots, posters, and metadata was developed by [5] who study the effectiveness of visual, audio, text, and metadata-based features for predicting high-level information about movies such as their genre or estimated budget.

The large-scale movie description challenge [9] was first held as a workshop at the International Conference on Computer Vision (ICCV) 2015. This was held as a unified challenge on Text generation using single video clip, and Text generation using single video clip as well as its surrounding context. This challenge was later also held at ICCV 2017. It was held as two challenges: The large scale movie description and understanding challenge [LSMDC] and [MovieQA]. The LSMDC task included Movie description, Movie annotation and retrieval, and Movie fill-in-the-blank task. The MovieQA task included Question-answering in movies and video retrieval based on plot synopses sentences. A follow-up challenge was also held at ICCV 2019.

Early visions of the proposed work [7] explored the usage of visual and audio descriptors, in addition to employing semantic analysis and linking with external knowledge sources in order to populate a knowledge graph. Under this approach, time-stamped multi-modal signals of the video would be uplifted into a format that could be utilised by the applications semantic model and the resulting knowledge graph could be searchable in a newly meaningful way.

ActivityNet [4] is a large scale video benchmark for human activity understanding. ActivityNet provides samples from 203 activity classes with an average of 137 untrimmed videos per class and 1.41 activity instances per video, for a total of 849 video hours. All ActivityNet videos are obtained from online video sharing websites. Amazon Mechanical Turk (AMT) workers are used to determine if videos contain the intended activity class. AMT workers also label the beginning and end points of each activity.

The How2 dataset [10] was introduced by Sanabria et al. This is a large scale multimodal language understanding dataset to aid in the development human like language understanding capabilities, where machines should be able to jointly process multimodal data, and not just text, images, or speech in isolation.

The TRECVID Instance Search (INS) [3] task is a benchmarking task run at NIST (U.S. National Institute of Standards and Technology) to measure systems performance on the retrieval of specific instances of either persons, locations, or objects. In the last three years the task targeted pairs of instances such as the retrieval of video shots of specific people in given locations within the British Broadcasting Corporation (BBC) Eastenders TV series. While recently the task targeted video shots of specific people doing specific actions in the Eastenders world.

TRECVID Video to Text (VTT) [3] is a benchmarking task also run at NIST to evaluate the performance of systems that automatically generate a single sentence description for short videos. Videos are typically shorter than 10 seconds in length and are taken from social media to represent real world situations. Although ActivityNet, INS and VTT benchmarks are all important efforts to encourage research in recognition and video semantics, they are still considered component tasks and not targeting the real holistic problem of high-level video understanding the way humans do.

3 TECHNICAL DETAILS

3.1 Dataset

The procurement of suitable datasets is vital for the undertaking of this new research area. The authors have spent time identifying Movies with a Creative Commons (CC) license [6] which can be disseminated to participating researchers for the purpose of this research. The most important criteria in selecting the movies were reasonable video quality, duration of more than 15 min at least, and self contained story lines with clear actors, relations, events and entities. In total, a pilot dataset of about 11 hrs has been collected from public websites such as Vimeo[1] and the Internet Archive[2].

Table 1 shows the current set of collected movies, their genre and durations. All movies have been deemed by the authors to be suitable for this research, and will be disseminated to participants as appropriate. The main challenge in this new task is related to the methods and techniques teams use and not the size of data set, We consider this data set to be large enough for the Grand Challenge and workshop in the first year of this task. The authors have also been in deliberations with the BBC regarding the licensing of the TV show *Land Girls* for use in this data set, and the licensing of this series has been approved and will be available for subsequent years of this task. This is a 3-season / 15-episode series set in World War 2 about the lives of a group of women doing their part for Britain in the *Women's Land Army* during the war.

Due to the limited number of movies available for this task, the above list will be split 50-50 between a development set and test set. Final ground truths for movies selected as the development set will be made available to participants. Participating researchers will

[1] https://vimeo.com/
[2] https://archive.org/

Movie	Genre	Duration
Honey	Romance	86 min
Let's Bring Back Sophie	Drama	50 min
Nuclear Family	Drama	28 min
Shooters	Drama	41 min
Spiritual Contact The Movie	Fantasy	66 min
Super Hero	Fantasy	18 min
The Adventures of Huckleberry Finn	Adventure	106 min
The Big Something	Comedy	101 min
Time Expired	Comedy / Drama	92 min
Valkaama	Adventure	93 min

Table 1: The HLVU Dataset of open source movies

Relationship	Inverse Relationship
Child of	Parent of
Spouse of	Spouse of
Sibling of	Sibling of
Descendant of	Ancestor of
Friend of	Friend of
In Relationship With	In Relationship With
Ambivalent Of	Is Not Liked By
Employee of	Employer of
Attends	Attended By
Colleague of	Colleague of
Apprentice of	Mentor of
Student of	Teacher of
Supervisor of	Subordinate of
Superintendent at	Responsibility of

Table 2: Example Relationships Between Characters

also be encouraged to procure their own suitable development sets and to share among other participating researchers.

3.2 Human Annotation Framework

A group of human assessors will be recruited to provide manual annotations to the above HLVU dataset. These annotations will be the basis from which queries and ground-truth are generated. Each of the movies listed above will be watched at least twice by annotators. The first time is to familiarize themselves with the movie story line and the main actors. In the second time watching the movie they will be asked to use the yEd [1] graphing tool to develop a knowledge graph encompassing all of the actors, entities, relationships, and events. An example knowledge graph developed using this software, modelled on *The Simpsons*, is shown in Figure 1. In this example Knowledge Graph we map out the relations between the main characters of the show, in the way that generated Knowledge Graph's should map out the relations between main characters of the movie used in the dataset for this task.

Annotators will be provided with a primary list of possible relations between characters from which they must choose the most accurate relationship. Table 2 provides some examples of the possible relations between characters and their inverse. If needed, they will be allowed to introduce new relationships as well.

In addition to the yEd graphing tool, a specially developed in-house annotation tool will also be used by annotators to save image captions of each of the actors and entities. This should also be used to list all the different actions which take place in the relationship between two people. Additionally this should be used to list key actions and events an individual has been involved in. These actions and events referred to here are not meant to be an exhaustive list and will be those actions and events that human annotators consider to be the key actions or events which may define the relationship between two individuals. In addition, the tool can be used to enable the annotators to document all the different events that happened during the movie and eventually enabling the development of a global ontology of events and actions relevant to the whole HLVU dataset. An example screen shot from using this annotation tool is shown in figure 2.

3.3 Query Development

The yEd graphing tool described in the previous section has the capability of exporting all entities and relationships to xgml or tgf format which can be used in the query development process to read all entities and relationships and develop different types of queries for evaluation purposes. The actions and events which may define the relationship between people may be derived from the in-house annotation tool described in the previous section as well. The following are a set of proposed query types which may be used to evaluate participating systems. Each query type tries to capture and test the systems for their comprehension of the tested movies from different points of view and levels of difficulty:

3.3.1 A- Fill in the graph space: Fill in spaces in the Knowledge Graph (KG). Given the listed relationships, events or actions for certain nodes, where some nodes are replaced by variables X, Y, etc., solve for X, Y etc. Example: **X** Married To Marge. **X** Friend Of Lenny. **Y** Volunteers at Church. **Y** Neighbor Of **X**. Solution for **X** and **Y** in that case would be: **X** = Homer, **Y** = Ned Flanders. A more formal query example may look like the following xml block (asking who is the spouse of "Marge"):

```
<Q.A>
  <Q.Id.1>
    <Subject>Person:Unknown_1</Subject>
    <Pred>Relation:Spouse_of</Pred>
    <Object>Person:Marge</Object>
  </Q.Id.1>
</Q.A>
```

3.3.2 B- Question Answering: This query type represents questions on the resulting KG, including actions and events, of the movies in the described dataset. For example, based on the Simpsons KG below, how many children does Marge have? By counting the 'Parent Of' edges from Marge to other nodes, we can see that Marge has two children, Bart and Lisa. Other possible questions may be for example: What does Ms. Krabappel do for a living? and the answer

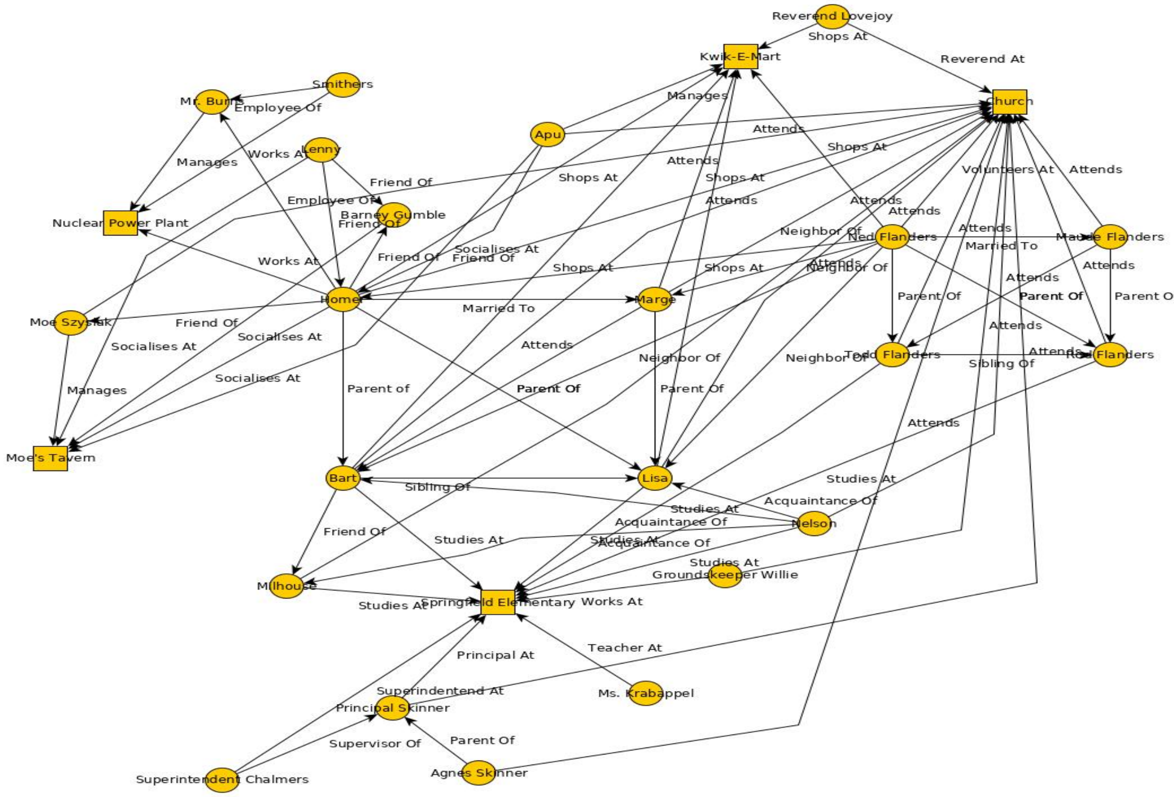


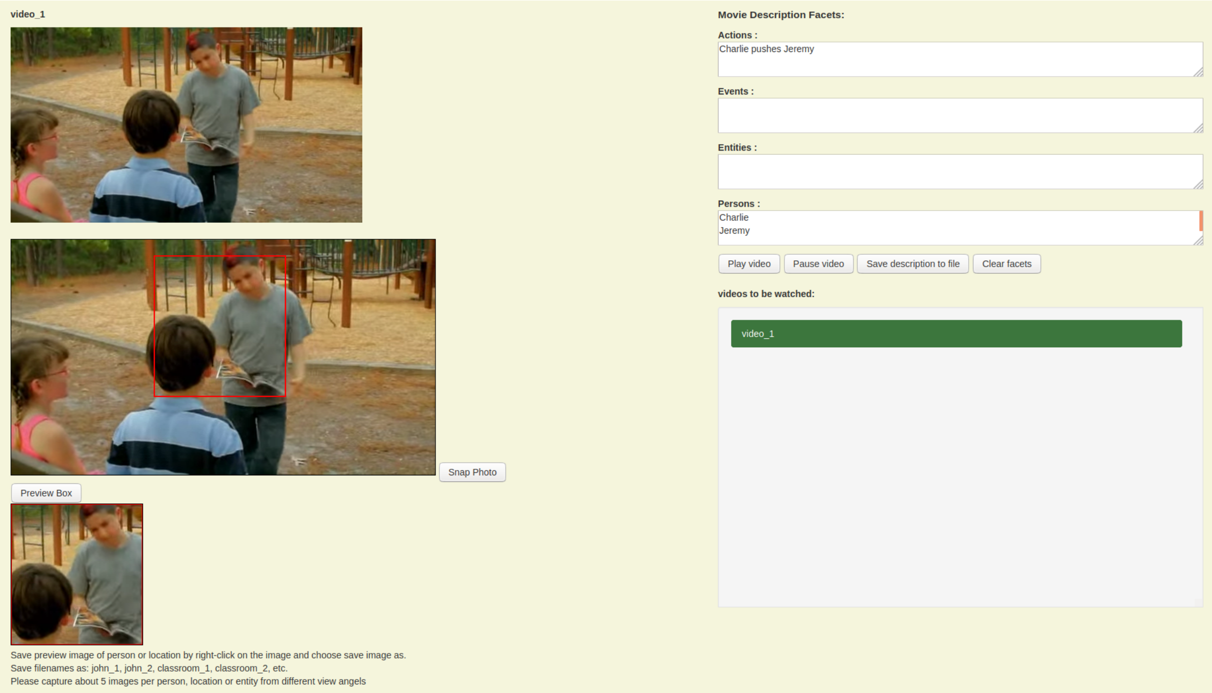
Figure 1: A sample knowledge graph depicting the world of *The Simpsons*

can be a multiple choice from the set of provided entity relationships. A query of this type may look like the following xml block:

```
<Q.B>
  <Q.Id.1>
    <Subject>Person:Ms. Krabappel</Subject>
    <Pred>Relation:Unknown_1</Pred>
    <Object>Location:Springfield Elementary</
        /Object>
    <Ans_1>Relation: X</Ans_1>
    <Ans_2>Relation: Y</Ans_2>
     .
     .
    <Ans_n>Relation: Z</Ans_n>
  </Q.Id.1>
</Q.B>
```

3.3.3 C- Relations between characters: How is character X related to character Y ? This query type question asks participants about all routes through the KG from one person to another. The main objective of this query type is to test the quality of the established KG. If the system managed to build a representative KG reflecting the real story line of the movie, then it should be able to return back all valid paths, including the shortest, between characters (i.e. how they are related to each other). For an example, looking at the Simpsons KG in figure 1, what would be the shortest route for Superintendent Chalmers, the left-bottom most node, to deliver a message to Lenny, in the top left hand corner of the knowledge graph? To answer this question we would follow the edges connecting each node and trace all possible paths from Superintendent Chalmers to Lenny, settling on the three shortest routes: (1) Superintendent Chalmers is supervisor to Principal Skinner, Principal Skinner attends Church, Church is also attended by Homer, Homer is in turn Lenny's friend. (2) Superintendent Chalmers is Superintendent at Springfield Elementary. Springfield Elementary is studied at by Bart. Bart is Child of Homer. Homer is friend of Lenny. (3) Superintendent Chalmers is Superintendent at Springfield Elementary. Springfield Elementary is studied at by Lisa. Lisa is Child of Homer. Homer is friend of Lenny. The below is the formal query representation for the above example:

```
<Q.C>
  <Q.Id.1>
    <Source>Person:Superintendent Chalmers</
        Source>
```



video_1

Movie Description Facets:

Actions :
Charlie pushes Jeremy

Events :

Entities :

Persons :
Charlie
Jeremy

Save preview image of person or location by right-click on the image and choose save image as.
Save filenames as: john_1, john_2, classroom_1, classroom_2, etc.
Please capture about 5 images per person, location or entity from different view angels

Figure 2: A screen shot of the annotation tool used for listing the defining actions and events for entities or relationships

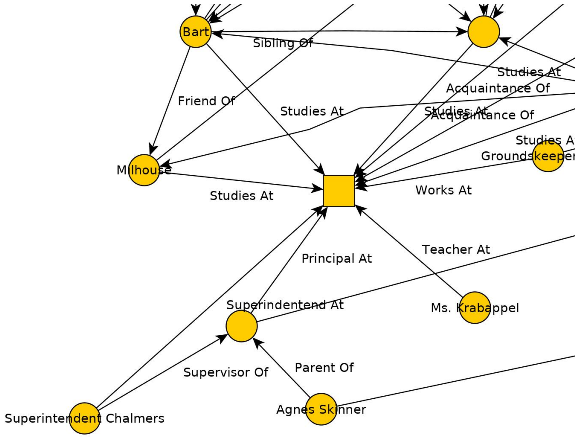


Figure 3: Visible region of graph with three blank nodes to be filled in.

```
    <Target>Person:Lenny</Target>
  </Q.Id.1>
</Q.C>
```

3.4 Evaluation Framework

Using the annotation and graphing tools described above, we can develop evaluation software to read as inputs the ground truth provided by these annotation and graphing tools, in addition to reading as input the submissions of participating teams, automatically evaluating each team's submissions, scoring and ranking results based on the selected evaluation metric. We should note here that systems will be given a set of image and/or video examples for the different actors and entities including important locations, each with a name Id. In addition, the final ontology of relationships, events, and actions will also be given so that systems can align the different query mentions of people, relations, entities, etc and return results from the given ontology classes as well. For illustration purposes, the following is a system response for each of the three query types and a proposed metric to score submissions:

3.4.1 A- Fill in the graph space: Given the partial graph query illustrated in figure 3, systems are asked to fill in the three blank spaces in the KG. The query itself will be given in xml format as illustrated in the query development section above. The ground truth for this example is: Lisa is sibling of Bart. Springfield Elementary is where Milhouse and Bart study at, and Principal Skinner is the principal. An example response by a system is shown below giving a set of answers for each unknown graph node being asked about in addition to a confidence score. Results will be treated as ranked list of result items per each unknown variable and the Reciprocal Rank score will be calculated per unknown variable and Mean Reciprocal Rank (MRR) per query [13].

```
<Q.A>
  <Q.Id.1>
    <Person:Unknown_1><Ans_1><Confidence>
    <Person:Unknown_1><Ans_2><Confidence>
     ...
     ...
    <Person:Unknown_2><Ans_n><Confidence>
  </Q.Id.1>
</Q.A>
```

The MRR measurement (Eq. 1) evaluates systems which return a ranked list of answers to questions. For this reason we consider it the most appropriate evaluation measurement for use in this query type. It is the average of the reciprocal ranks of results for a group of queries. For example, in three ranked lists submitted to answer three unknown variables in a query, if the answer to the first query is ranked second in the list, the answer to the second query is ranked first in the list and the answer to the third is ranked fourth, this gives $\frac{1}{2}$, 1, and $\frac{1}{4}$. This averages to give an MRR score of $\frac{7}{12}$.

$$MRR = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{RANKi} \tag{1}$$

3.4.2 B- Question Answering: Multiple choice questions will be provided for participants to answer which will be based on the KG. For an example on the format and types of questions asked to participants please see the Query Development section above. The below is an example of a system response and the evaluation metric (Eq. 2) proposed for this query type.

```
<Q.B>
  <Q.Id.1>
    <Answer>Relation: Z</Answer>
  </Q.Id.1>
</Q.B>
```

$$Score = \frac{CorrectAnswers}{TotalQuestions} \tag{2}$$

3.4.3 C- Relations between characters: In this query type, systems are asked to submit all valid paths form a source node to another target node with the goal of maximizing recall and precision. NIST will first evaluate whether each path is a valid path (i.e the submitted order of nodes and edges leads to a path from the source person to the target person of the query) and report the recall, precision and F1 measures [12]:

$$Recall = \frac{NumberofSubmittedValidPaths}{NumberofValidGTPaths} \tag{3}$$

$$Precision = \frac{NumberofSubmittedValidPaths}{NumberofSubmittedPaths} \tag{4}$$

$$F1\_Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

Below is an example of a system response to a query asking for all paths from the Simpsons characters Superintendent Chalmers to Lenny:

```
<Q.C>
  <Q.Id.1>
    <path=1>
    <Source>Person:Superintendent Chalmers</
        Source>
    <edge>Relation:Superintendent_At</edge>
    <node>Entity:Springfield Elementary</
        node>
    <edge>Relation:Studied_At_By</edge>
    <node>Person:Bart</node>
    .
    .
    <Target>Person:Lenny</Target>
    </path>
    <path=2>
    <Source>Person:Superintendent Chalmers</
        Source>
    .
    .
    </path>
    <path=3>
    .
    .
    </path>
  </Q.Id.1>
</Q.C>
```

The evaluation framework described above will provide a way to accurately and automatically evaluate the Knowledge Graphs produced by participating researchers. This allows for consistent scoring methods to be applied to this task which increases confidence in results and allows for easy expansion of the task and number of participants.

4 CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we discussed a proposed new evaluation benchmark and associated pilot dataset (HLVU) to promote research in holistic and deep video understanding. The evaluation paradigm aims for long duration self contained story lines in movies and targets automatic systems to eventually build complete knowledge graphs representing the movie characters, relationships, and key actions and events associated with them. The different query types presented aim to test the current automatic computer vision systems' capabilities and measure if they can "understand" video story lines the way humans do. The future plans for this evaluation benchmark is to organize an ACM Multimedia Grand Challenge and a workshop at ACM ICMI to host the evaluation campaign encouraging researchers to explore the dataset, discuss challenges and lessons

Curtis, Keith; Awad, George; Rajput, Shahzad; Soboroff, Ian. "HLVU: A new challenge to test deep understanding of movies the way humans do." Paper presented at ACM International Conference on Multimedia Retrieval (ICMR), Dublin, IE. June 08, 2020 - June 11, 2020.

learned trying to solve the associated problems with deep video understanding of the small world of movies. Also, securing more long term stable and large-scale dataset of licensed or open source movies including soap opera series of closed world story lines is a major future plan to continue measuring progress and the state of the art in this new domain.

ACKNOWLEDGMENTS

The authors would like to thank the Information Technology Laboratory at NIST for sponsoring the work presented in this paper.

Disclaimer: Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST, or the U.S. Government.

REFERENCES

[1] 2019. yEd - Graph Editor. https://www.yworks.com/products/yed, Last accessed on 2019-12-11.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[3] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, and Georges Quénot. 2019. TRECVID 2019: An evaluation campaign to benchmark Video Activity Detection, Video Captioning and Matching, and Video Search & retrieval. In *Proceedings of TRECVID 2019*. NIST, USA.

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 961–970.

[5] Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. 2019. Moviescope: Large-scale Analysis of Movies using Multiple Modalities. *arXiv preprint arXiv:1908.03180* (2019).

[6] Creative Commons. 2019. About The Licenses. https://creativecommons.org/licenses/, Last accessed on 2019-11-06.

[7] Jeremy Debattista, Fahim A Salim, Fasih Haider, Clare Conran, Owen Conlan, Keith Curtis, Wang Wei, Ademar Crotti Junior, and Declan O'Sullivan. 2018. Expressing Multimedia Content Using Semantics—A Vision. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE, 302–303.

[8] Bhavan Jasani, Rohit Girdhar, and Deva Ramanan. 2019. Are we asking the right questions in MovieQA?. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 0–0.

[9] Anna Rohrbach and Jae Sung Park. 2019. Large Scale Movie Description Challenge (LSMDC) 2019. https://sites.google.com/site/describingmovies/lsmdc-2019, Last accessed on 2019-11-06.

[10] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347* (2018).

[11] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4631–4640.

[12] Cornelis Joost Van Rijsbergen. 1979. Information retrieval. (1979).

[13] Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *Trec*, Vol. 99. Citeseer, 77–82.

Stress and Strain Heuristics for a Layered Elastomeric Foam at Impact Rates

Alexander K. Landauer, Jared Van Blitterswyk, Michael Riley, Aaron M. Forster

Materials Measurement Laboratory

National Institute of Standards and Technology

100 Bureau Drive, Gaithersburg, MD 20899

ABSTRACT

Impact mitigating materials (IMMs) are used to reduce injury or damage due to a blunt impact, which often occurs at high rates or energies. Innovation in IMMs and designs strategies are required for the development of safer protective equipment. A key challenge is translating between idealized experiments (e.g. quasistatic uniaxial stress or high rate Kolsky bar uniaxial strain measurements), fine-grain computational simulations, and real-world performance. To address this challenge, we have coupled high fidelity digital image correlation measurements with drop tower testing, based on our previous work presented to the SEM community. By using digital image correlation, an instrumented drop mass, and an instrumented load plate we obtain spatially and temporally resolved data for a realistic impact scenario. This represents an experimental framework that may be used to guide and validate design criteria applicable to the impact behavior of monolithic and layered IMMs.

Keywords: Impact protection, elastic foam, drop tower testing, digital image correlation, high-speed imaging

INTRODUCTION

Impact mitigating materials (IMMs) are designed and fabricated to minimize the effect of a mechanical insult or hit upon an object to be protected. Typical design targets include, e.g., minimizing linear and rotational accelerations, shock loading, energy transfer, specific or total mass, volume and/or cost, or maximizing energy dissipation [1, 2]. Two design paradigms exist: multi-use and single use. Single use applications can include plasticity or fracture-based solutions, whereas multi-hit applications are constrained to elastic or viscoelastic deformations that are recoverable over the expected time between impacts. Polymer foams, which consist of a crosslinked polymer matrix and gas or fluid filled negative space, are common in both paradigms due to the attractive and tailorable material properties, ease of manufacture and low cost, and adaptability due to the numerous chemistries and foaming processes available [1]. For example, rigid expanded polystyrene foams are ubiquitous for single-impact protection, whereas viscoelastic vinyl-nitrile foams are widely used for multi-impact protection. The uniaxial material response of a typical foam is represented in Fig. 1a, where the large compressive strain domain at relatively constant stress leads to a roughly constant load, i.e., an approximately constant linear acceleration of a protected mass, and the possibility for large energy dissipation if hysteretic mechanisms are active.

Here, we focus on specimens and test techniques for impact mitigation in personal protection, i.e., helmets, pads, or playing surfaces, in which multiple impacts over an O(1 s) time are expected. Typical examples include helmets and padding for contact sports (e.g., football, ice hockey, or lacrosse) and military and police armor and helmets, where impact are O(10 J) and O(10 ms) with strain rates in the padding material reaching 10^1 to 10^2 1/s. In practice, embodiments of protective systems across disciplines often have similar constructions that include a hard shell for penetration or fracture protection and one or more viscoelastic foam (or elastomeric) layers [3]. Foams are often arranged in series to provide both protection and comfort, e.g. a stiff foam adjacent to the exterior shell with an inner soft layer to interface with the head or body during normal use, see the schematic in Fig. 1b. A combination of material and structure are sought that minimize injury risk, either via top-down design, e.g., via models of discrete material layers that minimize stress wave propagation [4, 5] or promote dissipative mechanisms [3] or limit rotational (i.e., shear) loads [6], or via computational topology optimization of lattice-like architected structures [7, 8] which can then be fabricated through additive manufacturing techniques [9]. In this abstract, we discuss a combined instrumented drop tower and digital image correlation (DIC) system to interrogate the response of these structures in real-world impact-like conditions and show an example result for a layered foam.

Landauer, Alexander; Van Blitterswyk, Jared; Riley, Michael A.; Forster, Aaron M. "Stress and Strain Heuristics for a Layered Elastomeric Foam at Medium Impact Rates." Paper presented at Society for Experimental Mechanics, Bethel, CT, US. September 14, 2020 - September 17, 2020.

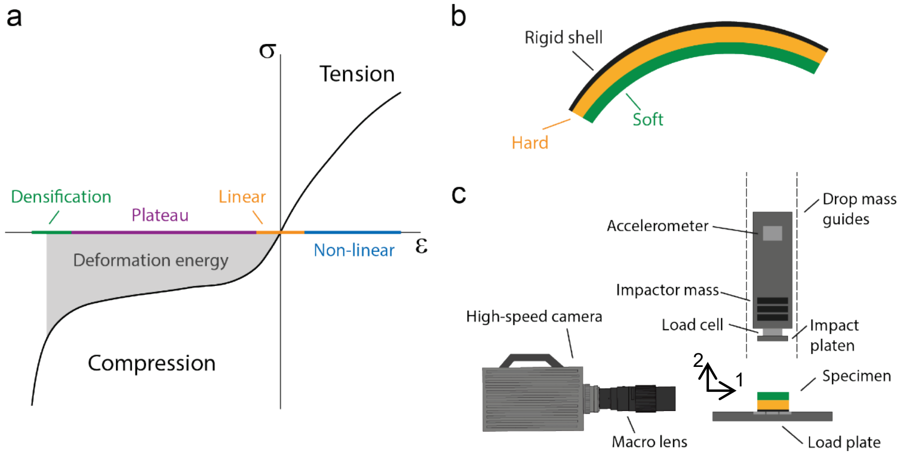


Fig. 1 (**a**) Typical stress-strain response of a foam-like material at constant strain rate. (**b**) Simplified schematic of a bi-layer IMM, with hard and soft layers of foam and a rigid external shell. (**c**) Components (not to scale) of the drop-mass system, including synchronized single camera and load-plate and instrumented (axial load and tri-axial acceleration) adjustable-weight drop mass

METHODS

Foam specimens approximately 50 mm x 50 mm x 25 mm were excised from as-received sheets of a stiff vinyl-nitrile foam (VNW, density approximately 111 kg/m^3) and a less stiff vinyl-nitrile foam (VNB, density approximately 135 kg/m^3). Both are closed-cell viscoelastic foams. Specimens were either monolithic foam or layered in a VNB (bottom) + VNW (top) or, as in the schematic of Fig. 1b and depicted in Fig. 3a, VNW (bottom) + VNB (top).

For our impact-like (i.e., O(10 J) and O(10 ms)) experiments we employ a custom-designed, instrumented drop-mass, which was previously presented to the SEM community [10]. A custom digital image correlation (DIC) algorithm [11] designed for finite deformations was incorporated into the instrumented drop tower experiments using high-speed imaging, see the schematic in Fig 1c. Typical settings and calibrations for DIC analysis are summarized in Table 1. Images for DIC were captured with a 1:2 macro lens (standoff distance approximately 0.5 m) at a framing rate of 10 kHz (exposure 1/20 000 s, f/5.6). The load plate captured synchronized force data at 50 kHz and the drop mass recorded triaxial accelerations and axial loads on an independent clock. The displacement field data is used to compute the surface strain tensor via a spatial differentiation filter. Strain rate history is computed using temporal central differencing. Table 2 describes the test conditions used in the drop tests. Monolithic materials were subjected to a 0.4 m drop of approximately 12 J of energy. Layered foams were approximately twice the height of the monolithic foams, therefore these were subjected to a 3x increase in energy, approximately 36 J, to achieve a similar degree of compressive strain as the monolithic cases.

RESULTS AND DISCUSSION

The scalar data outputs from the drop mass technique include the peak acceleration of the drop mass, the peak force applied to the top of the specimen, and peak force transmitted through the specimen. These results are given for the monolithic foams and both stack configurations in Table 2. The stiffer VNW was more effective at mitigating a 0.4 m drop (lower peak accelerations and forces) than the softer, but denser, VNB foam. Both foams had relatively short dominant relaxation times -

Table 1 Image and process parameters for digital image correlation of a typical monolithic specimen

Parameter	Value
Camera Noise	0.026%
Prefilter	Gaussian, 3 px by 3 px, σ = 0.5
Run mode	FFT-based, Hybrid stepping
Image size	818 px by 765 px
Subset size (final)	16 px by 16 px
Step	8 px
Strain filter	Optimal-7 tap
Virtual strain gauge	58 px
Temporal (rate) filter	Central difference
Interpolation	Cubic spline
Measurement points	10185
Total images	235
Static image displacement uncertainty	*x*: 0.020 px *y*: 0.019 px
Static image strain uncertainty	ϵ_{11} = 427 μm/m ϵ_{22} = 450 μm/m ϵ_{11} = 261 μm/m

Table 2 Drop mass impact data for impacts on monolithic and layered foams. Data are listed as mean ± one standard deviation for two repeats on two specimens (n = 4) for monolithic foam and two repeats per stack type for layered foam (n = 2)

Specimen	Drop Height (m)	Energy (J)	Peak impactor Acceleration (g)	Peak Applied Force (kN)	Peak Transmitted force (kN)
VNW	0.4	12.3	45.4 ± 3.5	1.09 ± 0.11	1.00 ± 0.07
VNB	0.4	12.3	63.0 ± 2.6	2.03 ± 0.07	1.96 ± 0.06
VNB+ VNW (top)	1.2	36.9	81.3 ± 0.3	2.73 ± 0.03	2.68 ± 0.12
VNW+ VNB (top)	1.2	36.9	71.2 ± 1.9	2.43 ± 0.06	2.31± 0.06

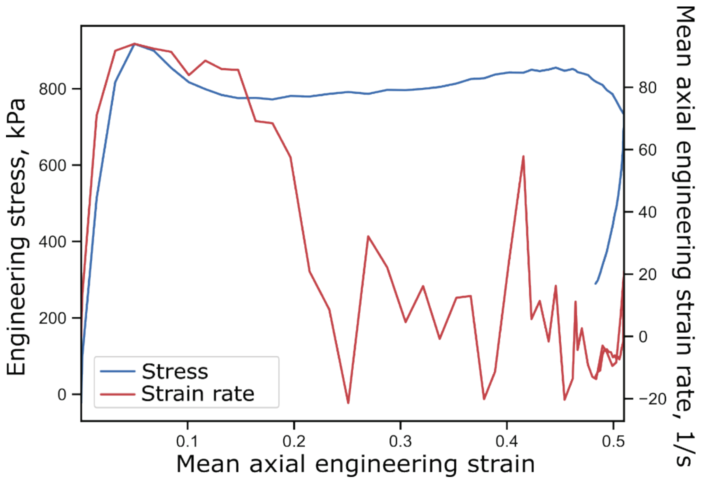


Fig. 2 A typical stress vs. strain and strain-rate vs strain result with mean axial engineering surface strain, ϵ_{22}, and mean axial engineering surface strain rate $\dot{\epsilon}_{22}$ from a 0.4 m impact on VNW

approximately 6.2% and 2.7% residual strain after 100 ms recovery time from impacts of approximately 36.9 J and 12.3 J yielding approximately 83% and 75% axial compressive engineering strain, respectively for VNW and VNB. For the monolithic foam samples, the axial engineering stress (left axis, computed via the measured cross section of the specimen) and the mean full-field axial strain rate (right axis) are plotted as functions of the mean full-field axial engineering strain, see Fig. 2. In this case, the 12.3 J impact leads to approximately 50% strain – enough to initiate an upturn in the stress response (i.e., densification of the foam), although the relative magnitude may be confounded by rate dependency in the material. Note, however, that this is semi-quantitative, since robust boundary conditions (i.e. uniaxial stress) are not established. In the initial loading the strain rate reaches approximately 90 1/s, but as the impact continues this is reduced to less than 20 1/s during the onset of densification.

In the multilayer case, a simple stress versus strain plot holds limited meaning, and instead we use a reduction of the DIC data to interpret the experiment. Peak acceleration and force measurements varied between the VNW+VNB and VNB+VNW configurations, indicating that the stacking order likely leads to measurable differences in impact mitigation. These effects can be further elucidated via the strain history from the DIC analysis. As an example, we take the case of a 36.9 J impact on a VNW+VNB layered foam, the undeformed configuration image of which is shown in Fig. 3a. Since the strain is highly non-uniformly distributed through the height of the specimen, a map of axial strain as a function of height of the specimen (the 2-direction) and post-impact time is given in Fig. 3b. Each time slice represents the mean axial engineering strain ϵ_{22} taken

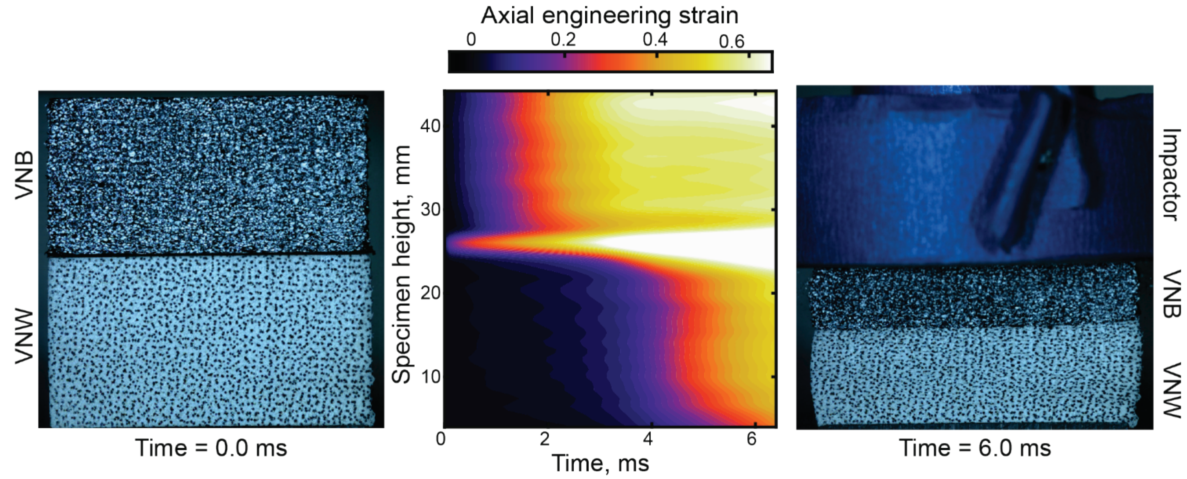


Fig. 3 Snapshot of the multilayer VNW+VNB foam in the undeformed and deformed configurations with the evolution of mean (in the 1-direction) axial strain ϵ_{22} during the first ca. 6 ms of a 36.9 J impact shown as a contour map

across the $x_1$direction. From this representation we see that in approximately the initial 2 ms of loading, the top layer deforms relatively uniformly until it reaches saturation at approximately 50% strain (i.e., approaching the densification strain). Apparent strain rapidly builds at the interface – likely an artificial strain since initial conformal contact is not guaranteed at the VNB/VNW interface and speckle pattern characteristic dramatically change – which warrants further investigation since interfaces are typically challenging for accurate DIC measurements. At approximately 4 ms the lower layer begins to rapidly take on strain, although less uniformly than the top layer, and this strain continues to increase beyond 6 ms into the impact event.

CONCLUSION

We have added digital image correlation to the instrumented drop tower experiment and used this to briefly investigate both monolithic and layered foams. The stress-strain and strain rate-strain behaviors, based on full-field measured strains of the specimen surface, are shown during the drop for monolithic foam. An example of layered foam with a strain distribution as a function of time is also given. These measurements highlight the complexity of the impact event compared to the well-controlled conditions of a conventional uniaxial compression test. The extraction of strains from a layered foam system highlights the non-uniformity of the material system response and possibly also the role of interfaces during impact. In the future, 3D-DIC will be employed to establish measurement criteria for elastomeric dissipative materials and the effects of layering strategies will be addressed.

ACKNOWLEDGEMENTS

The authors thank Dwight Barry for assistance fabricating foam specimens.

REFERENCES

[1] Mills N. Polymer Foams Handbook: Engineering and Biomechanics Applications and Design Guide. Elsevier, 2007.
[2] Gibson LJ, and Michael AF. Cellular Solids: Structure and Properties. Cambridge University Press, 1999.
[3] Duncan O. Shepherd T, Moroney C, Foster L, Venkatraman PD, Winwood K, Allen T, and Alderson A. “Auxetic foams for sports applications.” *Applied Sciences* 8, no. 6 (2018): 941. 10.3390/app8060941.
[4] Rahimzadeh T, Arruda EM, and Thouless MD. “Design of armor for protection against blast and impact.” *Journal of the Mechanics and Physics of Solids* 85 (2015): 98–111. 10.1016/j.jmps.2015.09.009.

[5] Sonnenschein MF, Edoardo N, Liangkai M, and Wendt BL. “Impact Mitigation in Layered Polymeric Structures.” Polymer 131 (2017): 25–33. 10.1016/j.polymer.2017.10.014.

[6] Johnston JM, Ning h, Kim JE, Kim YH, Soni B, Reynolds R, Cooper L, Andrews JB, and Vaidya U. “Simulation, fabrication and impact testing of a novel football helmet padding system that decreases rotational acceleration.” *Sports Engineering* 18, no. 1 (2015): 11–20. 10.1007/s12283-014-0160-4.

[7] Kuhn EN, Miller JH, Feltman B, Powers AK, Sicking D, and Johnston JM. “Youth helmet design in sports with repetitive low- and medium-energy impacts: a systematic review.” *Sports Engineering* 20, no. 1 (2017): 29–40. 10.1007/s12283-016-0215-9.

[8] Mueller J, and Shea K. “Stepwise graded struts for maximizing energy absorption in lattices.” *Extreme Mechanics Letters* 25 (2018): 7–15. 10.1016/j.eml.2018.10.006.

[9] Nazir A, Abate KM, Kumar A, and Jeng JY. “A state-of-the-art review on types, design, optimization, and additive manufacturing of cellular structures.” *The International Journal of Advanced Manufacturing Technology* 104, no. 9–12 (2019): 3489–3510. 10.1007/s00170-019-04085-3.

[10] Forster AM, Riley M. “Metrologies for Performance of Impact Mitigating Materials.” *SEM Annual Conference*, 2017. Indianapolis, IN, USA.

[11] Landauer AK, Patel M, Henann DL, and Franck C. “A q-Factor-Based Digital Image Correlation Algorithm (qDIC) for Resolving Finite Deformations with Degenerate Speckle Patterns.” *Experimental Mechanics* 58, no. 5 (2018): 815–30. 10.1007/s11340-018-0377-4.

Contrasting Conventional and Machine Learning Approaches to Optical Critical Dimension Measurements

Bryan M. Barnes*[1] and Mark-Alexander Henn[2]

[1]Nanoscale Device Characterization Division, Physical Measurement Laboratory
[2]Applied and Computational Mathematics Division, Information Technology Laboratory
National Institute of Standards and Technology 100 Bureau Drive MS 8423
Gaithersburg, MD USA 20899-8423

Abstract

Accurate, optics-based measurement of feature sizes at deep sub-wavelength dimensions has been conventionally challenged by improved manufacturing, including smaller linewidths, denser layouts, and greater materials complexity at near-atomic scales. Electromagnetic modeling is relied upon heavily for forward maps used to solve the inverse problem of optical measurements for parametric estimation. Machine learning (ML) approaches are continually under consideration, either as a means to bypass direct comparison to simulation or as a method to augment nonlinear regression. In this work, ML approaches are investigated using a well-characterized experimental data set and its simulation library that assumes a 2-D geometry. The benefits and limitations of ML for optical critical dimension (OCD) metrology are illustrated by comparing a straightforward library look-up method and two ML approaches, a data-driven surrogate model for nonlinear regression using radial basis functions (RBF) and multiple-output Gaussian process regression (GPR) that indirectly applies the simulated intensity data. Both RBF and GPR generally improve accuracy over the conventional method with as few as 32 training points. However, as measurement noise is decreased the uncertainties from RBF and GPR differ greatly as the GPR posterior estimate of the variance appears to overestimate parametric uncertainties. Both accuracy and uncertainty must be addressed in OCD while balancing simulation versus ML computational requirements.

1. INTRODUCTION

Inexpensive, non-destructive, and relatively fast optical approaches permeate overlay metrology,[1,2] defect inspection,[3,4] and critical dimension (CD) metrology[5] in advanced semiconductor manufacturing. Unlike defect and overlay metrologies, optical CD (OCD) metrology has been strongly dependent upon electromagnetic simulations to yield forward maps for the inverse problem of determining CDs and optical properties from intensity measurements. The downward scaling of semiconductor features has prompted not only improvements in electromagnetic simulations[6] but also reductions in measurement wavelength[7] and the rigorous incorporation of prior knowledge (e.g., hybrid metrology[8–12]) to extend the continued utility of OCD even as dimensions are deep sub-wavelength and approaching near-atomic scales.

The clear goal of OCD of sub-wavelength features is determining the set of parameters, $\mathbf{p} = \{p_1, \ldots p_T\}$, where T is the number of free parameters in the simulation, that best represent the measurand. Generally, optical data are collected for combinations of several measurement conditions (e.g., angle, wavelength, polarization, *etc.*) that can be labeled as $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$, where t is the total number of conditions. The optical data collected at each $\boldsymbol{\omega}_i$ can be classified as $\mathbf{I} = (I_1, \ldots, I_t)^\intercal$ and the simulated data are $\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) = [f(\boldsymbol{\omega}_1, \mathbf{p}), \ldots, f(\boldsymbol{\omega}_t, \mathbf{p})]^\intercal$. Most often, nonlinear regression is utilized to determine the estimate $\hat{\mathbf{p}}$ from the relationship

$$\mathbf{I} = \mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon}$ is an error term.

To improve OCD further, researchers and technologists are contemplating the use of artificial intelligence (AI) and machine learning (ML) towards determining $\hat{\mathbf{p}}$ with two divergent approaches emerging: First, one

*bmbarnes@nist.gov

Barnes, Bryan; Henn, Mark-Alexander. "Contrasting Conventional and Machine Learning Approaches to Optical Critical Dimension Measurements." Paper presented at Metrology, Inspection, and Process Control for Microlithography XXXIV, San Jose, CA, US. February 24, 2020 - February 27, 2020.

could abandon the long-held practice that experimental data need to be connected directly to simulation data. In simplest terms, the measurements $\mathbf{I}$ could be fed directly into ML to yield $\hat{\mathbf{p}}$. Second, ML could be centered on improving existing regression approaches for solving Eq. 1. In selecting between these two approaches for metrology, it is crucial that ML needs to yield not only $\hat{\mathbf{p}}$ but also the variances $\sigma^2_{\hat{\mathbf{p}}} = \{\sigma^2_{\hat{p}_1}, \ldots, \sigma^2_{\hat{p}_T}\}$. Of the first approach, Bischoff *et al.* reported fitting of 0.25 μm patterns using neural networks, but no uncertainties were reported.[13] Robert *et al.* reported an estimation of values using neural networks and also variance estimation using a second neutral network, but they clarify "variables are assumed to be not correlated". Rana *et al.* indicated industrial viability with reports of uncertainty using neural networks but without specific derivation.[14]

The second approach, the use of ML approaches to augment regression, has recently gained attention. Hammerschmidt *et al.* employed a Bayesian approach with a Newton-like method to quantify the parameters.[15] Heidenreich *et al.* reported a polynomial chaos based surrogate model for EUV scatterometry measurements with errors determined from Markov-Chain Monte Carlo sampling of the posterior distribution.[16] Later, Hammerschmidt *et al.* reported the use of Gaussian processes in Bayesian optimization, also referred to as Gaussian process regression in that work, using a Taylor expansion to quantify uncertainties in scatterometry.[17] In each of the works, results for the values were comparable with prior approaches at fitting.

In this work, previously published data are reprocessed to consider the challenges and opportunities of both the first and second approaches. The implementation of Gaussian process regression (GPR) presented here (unlike in Refs.[17,18]) circumvents direct application of Eq. 1 to proceed straight from $\mathbf{I}$ to $\hat{\mathbf{p}}$ through ML, although the ML training and validation are completely informed by rigorous electromagnetic simulations. GPR is a well-studied ML approach notable in that GPR can simultaneously calculate both a mean and a variance using multiple outputs (here, geometrical parameters) even with correlation.[19] Alternatively, radial basis functions (RBF) are trained and validated to augment the solution of Eq. 1 to yield improved values for the parametric means and variances.[20] These two approaches will be compared against a straightforward application of a conventional library look-up method. For the ML approaches, the complexity and computational expense of modern OCD modeling will be considered through the lens of data scarcity, a reduction in the number of training points, or n_{TP}, that are available in the existing library. Comparisons will show that even with smaller values of n_{TP}, both GPR and RBF yield generally good values for $\hat{\mathbf{p}}$, that RBF yields relatively small uncertainties based on its variance, but that this GPR method as presented appears to overestimate parametric uncertainties.

2. EXPERIMENTAL TECHNIQUE AND SAMPLE DETAILS

The experimental data used in this work were collected from scatterfield microscopy measurements in an angular scan mode. Scatterfield microscopy combines sophisticated illumination engineering in a high-magnification imaging platform with optimized information collection about targets of interest from the full 3-D electromagnetic scattered field. Overviews of the method are available as Ref.[22,23] Similar approaches have been deemed "micro-scatterometry".[24,25] Specifically, using Köhler illumination as illustrated in Fig. 1, an aperture in the conjugate to the back focal plane of the objective lens leads to an angularly resolved illumination beam (illumination NA

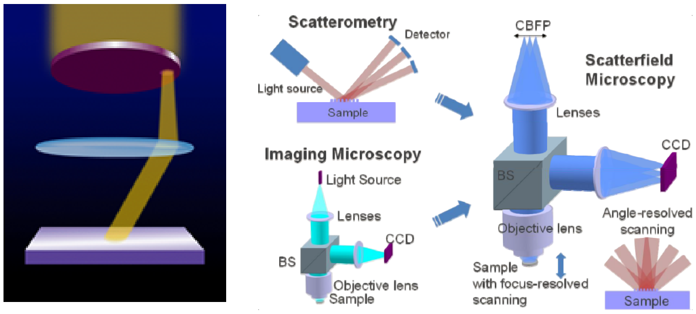


Figure 1. Angularly resolved scatterfield microscopy. (left) Schematic showing angularly resolved illumination. (right) Angle-scanning capabilities due to an aperture in the conjugate to the back focal plane (CBFP). Reprinted from Ref.[21]

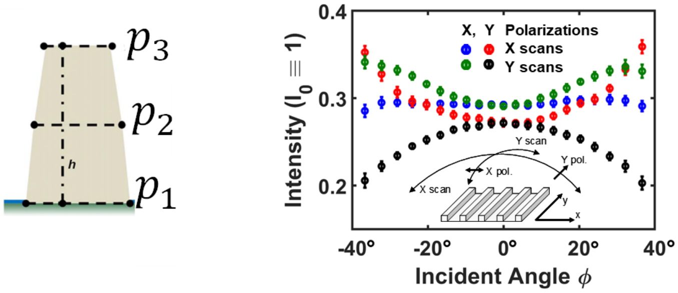


Figure 2. L100P300 Parameterized geometry and example data set. Measured intensities as functions of incident polarization and scan direction, with X, Y defined relative to line direction. Intensity here is unitless after normalization by the incident intensity $I_0 \equiv 1$. Error bars are 1σ uncertainties in the measured values. After Ref.[27]

$\approx$ 0.13) at the sample plane. This light is either reflected or scattered depending on the characteristics of the sample. The large collection NA $\approx$ 0.95 allows the capture of scattered light between $\phi = -72°$ to $72°$. For these experiments, the field-of-view of collection path was focused solely on the target of interest, a periodic array of patterned lines. The experimental wavelength is $\lambda = 450$ nm

The patterned structures under test are three scatterometry targets patterned using a focus/exposure matrix to yield variations in linewidth. The specific target is the "L100P300" target produced by SEMATECH,[26] with the notation implying a nominal line width of 100 nm and a period of 300 nm. These data have been of great utility, having been reported for initial demonstrations of Scatterfield Microscopy,[22] in pioneering work on hybrid metrology,[10, 11] as well as in a recent investigation of the treatment of potential experimental bias.[27] In these works, the geometry has been parameterized using a dual trapezoid model yielding three floating parameters, with a fixed height. Dimensional measurements from this sample have been reported from scanning electron microscopy (SEM)[22] and from atomic force microscopy (AFM).[8] Although its dimensions far exceed those of current technology nodes, comparisons can be made among the ML results as well as these prior data. Furthermore, with a large library of over 2000 simulated $\mathbf{f_I}(\boldsymbol{\omega}_i, \mathbf{p})$ in-hand, multiple realizations of sub-sampling of that library have been performed to yield additional rigor to the observed trends in accuracy and uncertainty.

3. METHOD DERIVATIONS

In this section, summaries describing the estimation of the parametric values $\hat{\mathbf{p}}$ and their uncertainties $\sigma_{\hat{\mathbf{p}}}$ are presented. Other sources cover the derivation of the RBF and GPR approaches more thoroughly (see the documentation for Ref.[20] and Ref.,[19] respectively).

3.1 Commonalities between Library look-up and RBF

Both library look-up and RBF utilize a weighted least squares fit. The weights are based on the known measurement errors, $\sigma_{\mathsf{meas},i}$, at each of the $\boldsymbol{\omega}_i, i = 1, 2, \cdots, t$ for the measurements I_i. In this work, $T = 84$ by concatenating the four angle scans in Fig. 2 into a single vector. This approach leads to the function

$$\chi^2(\mathbf{p}) = \sum_{i=1}^{t} \frac{1}{\sigma^2_{\mathsf{meas},i}} |f(\boldsymbol{\omega}_i, \mathbf{p}) - I_i|^2, \tag{2}$$

the weighted χ^2 (chi-square) function, which can also be expressed as:

$$\chi^2(\mathbf{p}) = [\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) - \mathbf{I}]^\intercal \, \mathbf{V}^{-1} \, [\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) - \mathbf{I}] \,, \tag{3}$$

with

$$\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p}) = [f(\boldsymbol{\omega}_1, \mathbf{p}), f(\boldsymbol{\omega}_2, \mathbf{p}), \ldots, f(\boldsymbol{\omega}_t, \mathbf{p})]^\intercal \,, \;\; \mathbf{I} = (I_1, I_2, \cdots, I_t)^\intercal \tag{4}$$

and for this work,

$$\mathbf{V}^{-1} = \left(\delta_{ij}\frac{1}{\sigma^2_{\mathsf{meas},i}}\right)_{i,j=1,\dots,t} \in \mathbb{R}^{t\times t}, \tag{5}$$

and as uncorrelated errors are assumed here $\mathbf{V}^{-1}$ is a diagonal matrix.

In addition, it can be shown that the uncertainties in $\hat{\mathbf{p}}$ can be calculated using the covariance matrix

$$\mathbf{\Sigma} = \left(\mathbf{J}^{\mathsf{T}}_{\mathbf{f_I}}\mathbf{V}^{-1}\mathbf{J}_{\mathbf{f_I}}\right)^{-1}, \tag{6}$$

where $\mathbf{J}_{\mathbf{f_I}}$ is the Jacobian of the model function. It is through Eqs. 5 and 6 that Library Look-up and RBF are both sensitive directly to measurement noise.

3.2 Library look-up

In the library look-up scheme, χ^2 is evaluated at a subset of $\mathbf{f_I}(\boldsymbol{\omega}_i, \mathbf{p})$ for which $\mathbf{p}$ are aligned on a grid such that for a given p_j there exists nearest neighbors (NN_-) and (NN_+) such that $p_{j(NN_-)} < p_j < p_{j(NN_+)}$ with $p_{i\neq j}$ being equal. The value of $\mathbf{p}$ which yields the smallest χ^2 is in this scheme the best fit $\hat{\mathbf{p}}$. The Jacobian is approximated using

$$\mathbf{J}_{\mathbf{f_I}} = \left(\frac{\partial f_i}{\partial p_j}\right)_{i=1,\dots,t,j=1,\dots,T} \in \mathbb{R}^{t\times T}, \frac{\partial f_i}{\partial p_j} \approx \frac{\Delta f_i}{\Delta p_j} \approx \frac{f_{i(NN_+)} - f_{i(NN_-)}}{p_{j(NN_+)} - p_{j(NN_-)}}. \tag{7}$$

Using additional nearest neighbors (if available) may improve this ratio through averaging of f_i, but here these simple approximations are used to quickly assess $\hat{\mathbf{p}}$ and $\sigma_{\hat{\mathbf{p}}}$.

3.3 Radial Basis Function Interpolation

Library look-up does not interpolate the function among points in the simulation domain, and for modern OCD such an extensive library seems impractical. One solution is to avoid evaluating the function directly by interpolating on a grid of already calculated values. Again, the RBFs are calculated using a subset of the entire library made up of n_{TP} entries where TP represents "training points"; this nomenclature comes from the ML community and shall be used for both RBF and GPR. The parameters for which $\mathbf{f_I}(\boldsymbol{\omega}, \mathbf{p})$ is evaluated are $\mathbf{p}_i = (p_{i1}, p_{i2}, \dots, p_{iT}), i = 1, \dots, n_{\mathsf{TP}}$ with

$$\mathbf{\Pi} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_{n_{\mathsf{TP}}} \end{pmatrix} \in \mathbb{R}^{n_{\mathsf{TP}}\times T}, \mathbf{f_I}(\mathbf{\Pi}) = \mathbf{\Phi} = \begin{pmatrix} \mathbf{f_I}(\mathbf{p}_1)^{\mathsf{T}} \\ \vdots \\ \mathbf{f_I}(\mathbf{p}_{n_{\mathsf{TP}}})^{\mathsf{T}} \end{pmatrix} \in \mathbb{R}^{n_{\mathsf{TP}}\times t}, \tag{8}$$

where T remains the number of parameters and t the number of measurement conditions, i.e., the length of an individual measurement vector.

The interpolation problem then amounts to find an approximation $\widetilde{\mathbf{f_I}}$ to the function $\mathbf{f_I}$ such that

$$\widetilde{\mathbf{f_I}}(\mathbf{p}, \mathbf{\Pi}, \mathbf{\Phi}) \approx \mathbf{f_I}(\mathbf{p}). \tag{9}$$

For the sake of clarity we will write $\widetilde{\mathbf{f_I}}(\mathbf{p})$ for the function in Eq. 9 if it is clear what library $\{\mathbf{\Pi}, \mathbf{\Phi}\}$ is being used.

Assume we have a set of functions $\rho_i : \mathbb{R}^T \to \mathbb{R},\ i = 1, \dots, N,$ such that for all $j \in \{1, \cdots, t\}$, we can approximate the j-th component of $\mathbf{f_I}$, i.e. the scalar function $f_j(\mathbf{p})$ by

$$f_j(\mathbf{p}) \approx \sum_{i=1}^{N} a_{ji}\rho_i(\mathbf{p}). \tag{10}$$

In the RBF approach specific functions are utilized that depend only on the distance of the parameter vector $\mathbf{p}$ from the different grid points $\mathbf{p}_i$, and an additional hyperparameter r as

$$\rho_i\left(\mathbf{p}, r\right) = \rho\left(\|\mathbf{p} - \mathbf{p}_i\|, r\right), \; i = 1, \ldots, n_{\mathsf{TP}}, \tag{11}$$

hence $N = n_{\mathsf{TP}}$ in Eq. 10. Specifically, the multiquadratic radial basis functions,

$$\rho_i\left(\mathbf{p}, r\right) = \sqrt{\|\mathbf{p} - \mathbf{p}_i\|^2 + r^2}, \; i = 1, \ldots, n_{\mathsf{TP}} \tag{12}$$

have been utilized with the hyperparameter r determined from leave-one-out cross-validation (LOOCV)[28] and the optimal hyperparameter $\hat{r}$ found using a particle-swarm optimization (PSO) algorithm, see Ref.[29] The matrix $\mathbf{A}$ (derived elsewhere[20]), consisting of the t coefficient vectors $\mathbf{a}_i = \left(a_{i1}, \ldots, a_{in_{\mathsf{TP}}}\right)^{\mathsf{T}} \in \mathbb{R}^{n_{\mathsf{TP}}}$ determined from Eq. 10, given as

$$\mathbf{A} = \left(\mathbf{a}_1, \cdots, \mathbf{a}_t\right)^{\mathsf{T}} \in \mathbb{R}^{t \times n_{\mathsf{TP}}}, \tag{13}$$

is also optimized using PSO. Using $\mathbf{A}$ we can calculate the approximation of all t entries to the model function for an arbitrary parameter vector $\mathbf{p}$ by

$$\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right) = \mathbf{A} \cdot \mathbf{P}\left(\mathbf{p}\right), \text{ with } \mathbf{P}\left(\mathbf{p}\right) = \left[\rho_1\left(\mathbf{p}\right), \ldots, \rho_{n_{\mathsf{TP}}}\left(\mathbf{p}\right)\right]^{\mathsf{T}} \in \mathbb{R}^{n_{\mathsf{TP}}}. \tag{14}$$

To quickly summarize, the training step for RBF solves for $\hat{r}$ and $\mathbf{A}$ that are used in the determination of $\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right)$ which can be shown to yield

$$\hat{\mathbf{p}} = \operatorname{argmin}\left\{\left[\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right) - \mathbf{y}\right]^{\mathsf{T}} \mathbf{V}^{-1} \left[\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right) - \mathbf{y}\right]\right\}, \tag{15}$$

where $\mathbf{y}$ is a general form, and is $\mathbf{I}$ specifically here as in Eq. 3. The derivation of the Jacobian of $\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right)$, $\mathbf{J}_{\widetilde{\mathbf{f}}_{\mathbf{I}}}$, can be found elsewhere,[20] but as with the library look-up, the parametric mean and uncertainty are both directly tied to the measurement uncertainty through $\mathbf{V}^{-1}$.

3.4 Gaussian Process Regression

In this implementation of Gaussian process regression, which has followed the derivation set forth by Liu *et al.*,[19] we rethink the relationships between the geometry and its scattering and attempt a vastly different ML approach. Instead of trying to establish a functional relationship between the geometry parameters and the optical response, i.e., trying to find the function

$$\mathbf{f}_{\mathbf{I}} : \mathbf{p} \mapsto \mathbf{f}_{\mathbf{I}}\left(\mathbf{p}\right), \tag{16}$$

and minimize the difference between a particular measurement $\mathbf{y}$ and the function value to determine the optimal value of geometry parameters, we want to establish a direct relationship between measured intensities and geometry parameters, i.e., find a function such that

$$\mathbf{g} : \mathbf{I} \mapsto \mathbf{g}_{\mathbf{p}}\left(\mathbf{I}\right), \tag{17}$$

informed by observed measurements and/or simulated values $\mathbf{I}$ at various measurement conditions $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$.

In order to keep the notation simple we will in the following assume that we only want to determine a single parameter value p instead of a vector of parameter values $\mathbf{p}$, before giving a brief explanation on how to address the latter with a multi-output Gaussian process (MOGP). Similar to the previous approach we assume that the observations, that are now the values of the parameter of interest, are noisy realizations of the underlying model, such that

$$p = g_p\left(\mathbf{I}\right) + \epsilon_s, \; \epsilon_s \sim \mathcal{N}\left(0, \sigma_s^2\right). \tag{18}$$

Note, that there is no relationship between ϵ_s in Eq. 18 and ϵ in Eq. 1, nor is there a relationship between σ_s and the measurement noise σ_{meas} in Eq. 2.

In Gaussian process regression the target function in Eq. 18 is approximated by interpreting it as a probability distribution in a function space, more precisely as a collection of random variables, such that any finite number of which have a joint Gaussian distribution. As such it is completely defined by its mean function $m(\mathbf{I})$ and the covariance function $k(\mathbf{I},\mathbf{I}')$. We can therefore write a Gaussian process as

$$g_p(\mathbf{I}) \sim \mathcal{GP}\left[m(\mathbf{I}), k(\mathbf{I},\mathbf{I}')\right], \tag{19}$$

and furthermore, w.l.o.g., assume the mean function to be zero. In order to reduce the complexity of the problem the choice of the covariance function is limited to a certain class of functions that can be characterized by a few parameters.

In this study, a simple exponential (SE) kernel has been applied, defined here as

$$k_{\mathsf{SE}}(\mathbf{I},\mathbf{I}') = \sigma^2 \mathsf{exp}\left(-\frac{||\mathbf{I}-\mathbf{I}'||}{2l}\right), \tag{20}$$

where the signal variance σ^2 represents an output scale amplitude and l represents a characteristic length scale. Here, σ and l are the hyperparameters ω for this kernel that must be fitted in order to realize the optimal GPR. Indeed, not only is this similar to RBF, but this kernel is itself a radial basis function also.

Because of the Euclidian norm in Eq. 20, a function of the vector difference between two intensity vectors with t elements is reduced to a scalar value. This is both a benefit (by speeding up the calculation) and a challenge (by removing information from the regression). Contrast this with the RBF interpolation, which in Eqn. 11 reduces a function of two vectors of length n_{TP} into a scalar value also. Alternative kernels including non-radial kernels should be explored further for use in GPR-based OCD metrology.

Thus, in the context of Eq. 17, there exists a set of training points $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_{n_{\mathsf{TP}}}\}$ from the input domain with associated output observations $\mathcal{P} = \{p(\mathbf{I}_1), \ldots, p(\mathbf{I}_{n_{\mathsf{TP}}})\}$, that we can arrange in a vector $\mathbf{p} = [p(\mathbf{I}_1), \ldots, p(\mathbf{I}_{n_{\mathsf{TP}}})]^{\mathsf{T}}$. Assume now that $\mathbf{I}$ need not (as defined above) be from simulation but rather comes instead from experiment as $\mathbf{y}$. Again following Ref.[19] and since a Gaussian process is a stochastic process wherein a finite subset of random variables follows a joint Gaussian distribution, the joint prior distribution of the observations $\mathbf{p}$ together with $g_p(\mathbf{y})$ is

$$\begin{bmatrix} \mathbf{p} \\ g_p(\mathbf{y}) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\mathcal{I},\mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}} & \mathbf{k}(\mathcal{I},\mathbf{y}) \\ \mathbf{k}(\mathbf{y},\mathcal{I}) & k(\mathbf{y},\mathbf{y}) \end{bmatrix}\right), \tag{21}$$

where $\mathbf{K}(\mathcal{I},\mathcal{I}) \in \mathbb{R}^{n_{\mathsf{TP}} \times n_{\mathsf{TP}}}$ is the symmetric and positive semi-definite covariance matrix with the element $\mathbf{K}_{ij} = k(\mathbf{I}_i, \mathbf{I}_j)$, $\mathbf{k}(\mathcal{I},\mathbf{y}) \in \mathbb{R}^{n_{\mathsf{TP}}}$ denotes the vector of covariances between the n_{TP} training points and the test point $\mathbf{y}$, and $\mathbf{E}_{n_{\mathsf{TP}}}$ is the identity matrix in $\mathbb{R}^{n_{\mathsf{TP}} \times n_{\mathsf{TP}}}$. Be aware that we understand the term point in a topological sense, since both the training points and the test point are vector quantities.

It can be shown that for a single-output Gaussian process regression, the prediction mean $\hat{g}_p(\mathbf{y})$ and prediction variance $\sigma_p^2(\mathbf{y})$ are

$$\hat{g}_p(\mathbf{y}) = \mathbf{K}(\mathbf{y},\mathcal{I})^{\mathsf{T}} \left[\mathbf{K}(\mathcal{I},\mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}\right]^{-1} \mathbf{p}, \tag{22}$$

and

$$\sigma_p^2(\mathbf{y}) = k(\mathbf{y},\mathbf{y}) - \mathbf{K}(\mathbf{y},\mathcal{I})^{\mathsf{T}} \left[\mathbf{K}(\mathcal{I},\mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}\right]^{-1} \mathbf{K}(\mathbf{y},\mathcal{I}). \tag{23}$$

To use Eqs. 22 and 23 for prediction, we need to infer the hyperparameters $\boldsymbol{\theta}$ in the covariance function k by minimizing the negative log marginal likelihood (NLML) as

$$\boldsymbol{\theta}_{\mathsf{opt}} = \arg_{\boldsymbol{\theta}}\min\ \mathsf{NLML}, \tag{24}$$

where

$$\mathsf{NLML} = -\log \pi(\mathbf{p}|\mathcal{I},\boldsymbol{\theta}) = \frac{1}{2}\mathbf{p}^{\mathsf{T}}[\mathbf{K}(\mathcal{I},\mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}]^{-1}\mathbf{p} + \frac{1}{2}\log|\mathbf{K}(\mathcal{I},\mathcal{I}) + \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}}| + \frac{n_{\mathrm{TP}}}{2}\log 2\pi. \tag{25}$$

After inferring $\boldsymbol{\theta}$, GPR can be applied to test data sets.

For brevity, we also state without additional derivation the formulation of a multi-output GPR in which we want to determine a T-dimensional parameter vector. We will assume that we have observations of the multiple parameters that share the same underlying set of training points, i.e., we have one set $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_{n_{\mathsf{TP}}}\}$ from the input domain and n sets of corresponding output observations $\mathcal{P}_i = \{p_i(\mathbf{I}_1), \ldots, p_i(\mathbf{I}_{n_{\mathsf{TP}}})\}$, each of which organized in a vector $\mathbf{p}_i = [p_i(\mathbf{I}_1), \ldots, p_i(\mathbf{I}_{n_{\mathsf{TP}}})]^\intercal$, and define the concatenated vector $\mathbf{p} = [\mathbf{p}_1^\intercal, \ldots, \mathbf{p}_T^\intercal]^\intercal$. For the MOGP we will furthermore assume that we have a single kernel function k that accounts for the different parameters by different coefficients a_{ii} for the different output parameters. Furthermore using a single underlying kernel function allows us to take correlations between the different output parameters into consideration by defining:

$$\mathbf{K}_M(\mathcal{I},\mathcal{I}) = \mathbf{A} \otimes \mathbf{K}(\mathcal{I},\mathcal{I}) = \begin{bmatrix} a_{11} \cdot \mathbf{K}(\mathcal{I},\mathcal{I}) & \cdots & a_{1T} \cdot \mathbf{K}(\mathcal{I},\mathcal{I}) \\ \vdots & \ddots & \vdots \\ a_{T1} \cdot \mathbf{K}(\mathcal{I},\mathcal{I}) & \cdots & a_{TT} \cdot \mathbf{K}(\mathcal{I},\mathcal{I}) \end{bmatrix} \in \mathbb{R}^{n_{\mathsf{TP}} \cdot T \times n_{\mathsf{TP}} \cdot T}, \tag{26}$$

with $\mathbf{K}(\mathcal{I},\mathcal{I})$ as in the single-output GPR. In this case the prediction mean and variance are given as

$$\hat{g}_{\mathbf{p}}(\mathbf{y}) = \mathbf{K}_M(\mathbf{y},\mathcal{I})^\intercal \left[\mathbf{K}_M(\mathcal{I},\mathcal{I}) + \boldsymbol{\Sigma}_s\right]^{-1} \mathbf{p}, \tag{27}$$

and

$$\boldsymbol{\Sigma}(\mathbf{y}) = \mathbf{K}_M(\mathbf{y},\mathbf{y}) - \mathbf{K}_M(\mathbf{y},\mathcal{I})^\intercal \left[\mathbf{K}_M(\mathcal{I},\mathcal{I}) + \boldsymbol{\Sigma}_s\right]^{-1} \mathbf{K}_M(\mathbf{y},\mathcal{I}), \tag{28}$$

respectively. Here

$$\boldsymbol{\Sigma}_s = \sigma_s^2 \mathbf{E}_{n_{\mathsf{TP}}} \otimes \mathbf{E}_T \in \mathbb{R}^{n_{\mathsf{TP}} \cdot T \times n_{\mathsf{TP}} \cdot T},$$

and $\mathbf{K}_M(\mathbf{y},\mathbf{y})$ and $\mathbf{K}_M(\mathbf{y},\mathcal{I})$ are defined analogous to the single-output case taking into account Eq. 26. Note that in the learning phase of the MOGP we therefore need to determine the entries $\{a_{ij}\}_{i,j=1,\ldots,T}$ of the matrix $\mathbf{A}$ in addition to the kernel's hyperparameters $\boldsymbol{\theta}$. However, requiring the matrix to be symmetric and positive semi-definite helps to reduce the total number of floated variables.

4. TRAINING AND VALIDATION DATA

4.1 Methodology

For the two ML approaches, RBF and GPR, the simulation data will be treated as if these data are computationally expensive, while utilizing over 1000 library points on a grid in library look-up as a comparison. For the ML approaches, this data scarcity will appear similar to Fig. 3, except that at each n_{TP} the actual points in the parameter space will be varied at each realization. Specifically, for $n_{\mathsf{TP}} = 16, 32, \ldots, 128$ there are 112 realizations for each of five values of the simulated "measurement" noise $\sigma_{\mathsf{noise}} = 10^i, i = \{-6, \ldots, -2\}$ where the incident intensity $I_0 \equiv 1$. For $n_{\mathsf{TP}} = 256$, the number of realizations is smaller, between 64 and 80. In training and validation, $\sigma_{\mathsf{meas}} \equiv \sigma_{\mathsf{noise}}$. For all methods, this noise is applied to the validation data while for RBF and library-look up, it is also used in the formation of the matrix $\mathbf{V}^{-1}$.

As can be inferred from the small markers in Fig. 3, the simulation data library is not equally distributed throughout the simulation domain. In each realization, the validation set is indexed randomly while the n_{TP}

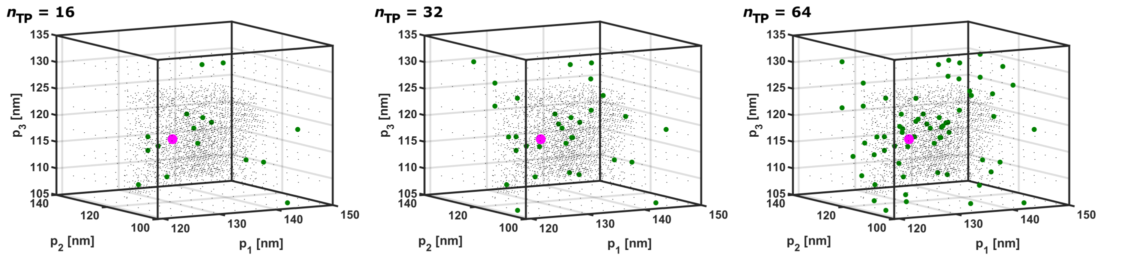


Figure 3. Examples of data scarcity in the simulation space. Validation, training points vary among n_{TP} except in Sec. 4.4
.

training data are selected through draws of random values for the parameters, normally distributed about the center of the simulation domain. With each triplet of values drawn, the closest library entry is indexed. If the candidate index matches the validation point or is repeated, a new candidate is drawn until n_{TP} points are determined.

Two key metrics are required to evaluate these approaches. It is clear that the parametric uncertainties of each parameter must be compared against each other, as well as the parametric values. A single value for comparing the accuracy among the methods can be taken from the GPR literature,[19] the relative average absolute error (RAAE). The RAAE is defined for a general function $f(\mathbf{x})$ and its expectation value $\hat{f}$ as

$$\text{RAAE} = \frac{\Sigma_{i=1}^{n_{\mathsf{TP}}} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|}{n_{\mathsf{TP}} \times \sigma(f_{\mathsf{train}}(\mathbf{x}))}, \tag{29}$$

where $\sigma(f_{\mathsf{train}}(\mathbf{x}))$ is the standard deviation of the training data.

In the analysis that follows, the RAAE and parametric uncertainties are compared after hundreds of realizations. Fig. 4 illustrates one such metric, the uncertainty in p_2 computed using all entries in the library look-up at $\sigma_{\mathsf{noise}} = 10^{-3}$. To compare among the different approaches, bar and whisker plots will illustrate the distribution of the RAAE and uncertainty from the training and validation steps and the parametric values from RBF and GPR in the testing step.

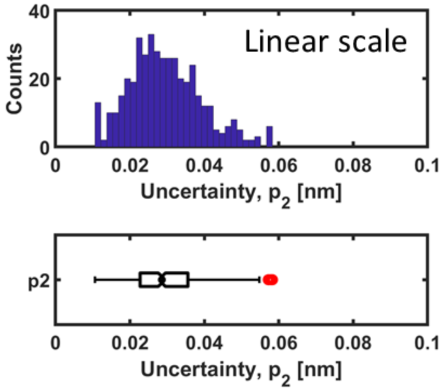


Figure 4. (top) Histogram of uncertainty values from library look-up for $\sigma_{\mathsf{noise}} = 10^{-3}$. (bottom) Example box-and-whisker plot for quickly conveying the distribution from these multiple realizations. In this work, lower and upper box edges show the 25th and 75th percentiles respectively (q_{25}, q_{75}); narrowest notch point is the median (50th percentile); minimum whisker length extends to the outermost data point beyond the box (see left side) while maximum whisker length is $1.5 \times (q_{75} - q_{25})$ (see right side); each outlier beyond the whisker range is shown individually using a symbol.

4.2 Accuracy

The accuracy of the library look-up mode is illustrated at the bottom of in Fig. 4. These data are to be compared against the resulting RAAE values from RBF and GPR as functions of n_{TP} with $\sigma_{\mathsf{noise}} = 10^{-3}$.

Figure 5 shows that for this three-parameter model and sample, only $n_{\mathsf{TP}} \geq 32$ is required to attain a lower (thus better) RAAE than from library look-up. Both the augmenting of the regression through RBF and the indirect use of simulation data through GPR allow far fewer simulations for improved validation set accuracy over the more straightforward conventional approach. Accuracy as expected improves with increases in n_{TP}. Values for the RAAE between GPR and RBF are similar.

4.3 Uncertainty

The major difference observed in this investigation is among the uncertainties in the validation set. As seen in Section 3, both the estimated parametric mean and parametric uncertainties can be determined from RBF and GPR without disregarding correlations among parameters.

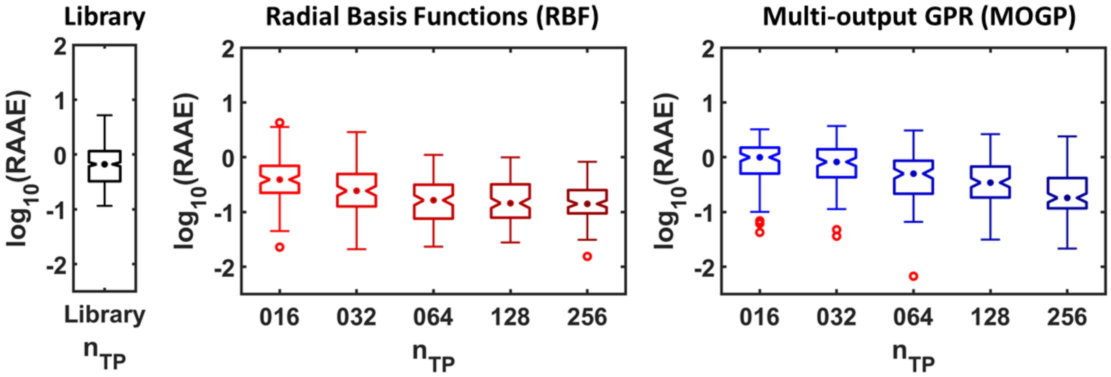


Figure 5. Accuracy of RBF and GPR versus library look-up as a function of n_{TP} for these training and validation data. In general, increasing n_{TP} decreases RAAE, indicating improved accuracy. If notches between any two of these box plot do not overlap, one may conclude that their two medians differ with 95 % confidence.

A key test is the scaling of the uncertainty with the measurement noise. As also established in Section 3, the measurement uncertainty is incorporated into the covariance matrix for both library look-up and RBF such that the parametric uncertainties should scale linearly with σ_{noise}. However, σ_{noise} only indirectly enters into the estimation of the covariance matrix through the application of this noise to the ”measurement” data. It was uncertain if this would translate into changes in uncertainty.

In Fig. 6, the expected behavior is observed from library look-up and RBF. Furthermore, the RBF interpolation consistently appears to reduce the uncertainty relative to library look-up by about two orders of magnitude. However, it is unclear if the measurement noise affects the GPR similarly. Here, there is a slight uptick in uncertainty between $\sigma_{\mathsf{noise}} = 10^{-3}$ and $\sigma_{\mathsf{noise}} = 10^{-2}$, but the uncertainty remains unchanged for $\sigma_{\mathsf{noise}} \leq 10^{-3}$. Notably, GPR yields a higher estimate of the uncertainty than either library look-up or RBF interpolation, suggesting a near constant factor in the uncertainty that is independent of measurement noise except for large values of σ_{noise}. Note also, training of the RBF involves solving for both $\mathbf{A}$ in Eq. 13 and hyperparameter r from Eqn. 11, a total of $1 + (t \times n_{\mathsf{TP}})$ parameters, while here GPR training solves for the two hyperparameters σ and l in Eqn. 20. Incorporating additional hyperparameters in the GPR might positively influence the uncertainty from GPR.

4.4 Monte Carlo Assessment of GPR Uncertainty

To assess this nature of the uncertainty estimated from GPR, the numerical experiment was repeated for the multi-output GPR, but instead of picking a new subset of points from the library of size n_{TP} on each realization,

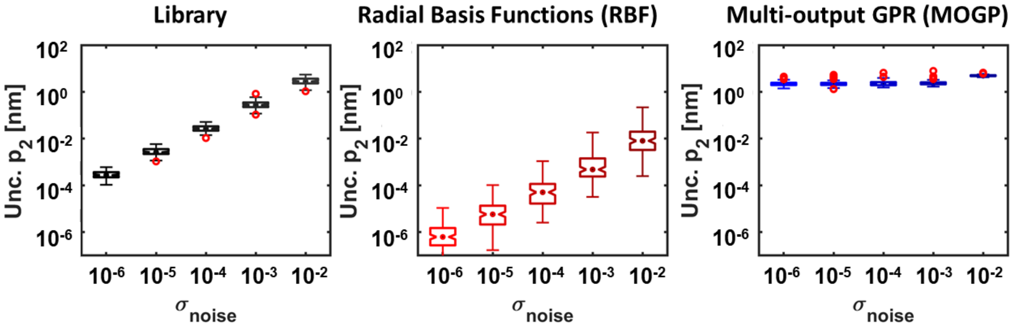


Figure 6. Uncertainty of parameter 2 (middle width) of RBF and GPR versus library look-up as a function of σ_{noise} for these training and validation data. As intensities have been normalized to the incident intensity, σ_{noise} is also unitless.

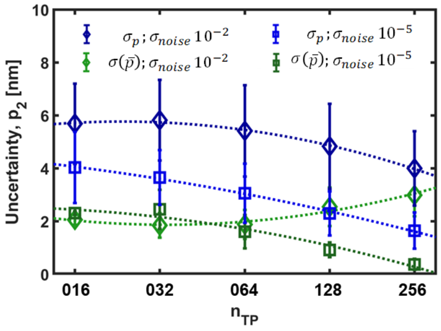


Figure 7. Comparisons of the variance of the middle width estimated by the GPR, σ_p^2, to the variance of the middle width $\sigma^2(\bar{\mathbf{p}})$ from Monte Carlo as a function of n_{TP}, the number of training points. Error bars are one standard deviation from 112 realizations.

multiple realizations were performed using the exact same points at each realization for n_{TP} points. Furthermore, as n_{TP} increased, the previous points were retained. That is, there were not new random draws for $n_{\mathsf{TP}} = 64$ compared to $n_{\mathsf{TP}} = 32$ but rather the subset for $n_{\mathsf{TP}} = 64$ contained 32 new points as well as the points considered at $n_{\mathsf{TP}} = 32$ (which also contained the points for $n_{\mathsf{TP}} = 16$.) For all these simulations, the validation point also remained the same with only its measurement noise varied on each realization, applied at each of the experimental measurement conditions $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$ independently assuming a Gaussian distribution. GPR was performed for 112 realizations for each combination of n_{TP} and σ_{noise}. This approach not only allows the estimation of the posterior mean with its variance, $\sigma(\mathbf{p})$, but also allows computation of the variance of the mean parametric value $\bar{\mathbf{p}}$, or $\sigma^2_{\bar{\mathbf{p}}}$. The uncertainties are shown in Fig. 7 for $\sigma_{\mathsf{noise}} = 10^{-2}$, a relatively large amount of measurement noise and also a much lower value, $\sigma_{\mathsf{noise}} = 10^{-5}$.

Figure 7 illustrates a nearly factor of two difference between the uncertainties $\sigma(\bar{\mathbf{p}})$ and σ_p for low noise. From this additional investigation, GPR as presented here may overestimate its variance except for high noise and a relatively large number of training points. Note that the mean of the parameter estimate only approaches zero for low noise and $n_{\mathsf{TP}} = 256$. Additional work is required to further close the two orders-of-magnitude or more gap in uncertainties between RBF and GPR observed in Fig. 6.

5. RBF AND GPR TESTING

With the discrepancies among RBF and GPR uncertainties, only the parametric values will be considered from these testing results. The experimental measurement noise varied for each measurement condition $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_t$ but on average, with the incident intensity $I_0 \equiv 1$, the magnitude fell between $\sigma_{\mathsf{noise}} = 10^{-2}$ and $\sigma_{\mathsf{noise}} = 10^{-3}$, for reference.

In Ref.[8] three specific target dies were measured and reported. For consistency, these same three die are measured here using the library look-up and ML approaches. In Fig. 8, the library look-up value is plotted as the dotted line in each individual panel. The leftmost column shows the measurement from each of the die and its label as identified in Ref.:[30] Dies (-1,-1), (0,0), and (1,1). The top row for each die shows the GPR fits to the three parameters identified in Fig. 2, while the bottom row for each die shows the RBF results. For Dies (1,1) and (0,0), the ML-based fits generally agree with the conventional library look-up approach.

Notable problems appear for the RBF fits for the third die, Die (-1,-1). Noted with ellipses on the bottom-most row, there is strong evidence of the distribution of parametric value splitting into two values across the range of n_{TP}. This is highly indicative of an additional local minimum to the weighted χ^2 fitting. Fortunately, the literature has shown a path towards a remedy in two steps. First, in the documentation to Ref.,[20] it is clear that the optimal parameter from RBF can be evaluated using and

$$\hat{\mathbf{p}} = \operatorname{argmin}\left[\left(\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right) - \mathbf{y}\right)^{\mathsf{T}} \mathbf{V}^{-1} \left(\widetilde{\mathbf{f}}_{\mathbf{I}}\left(\mathbf{p}\right) - \mathbf{y}\right)\right]. \tag{30}$$

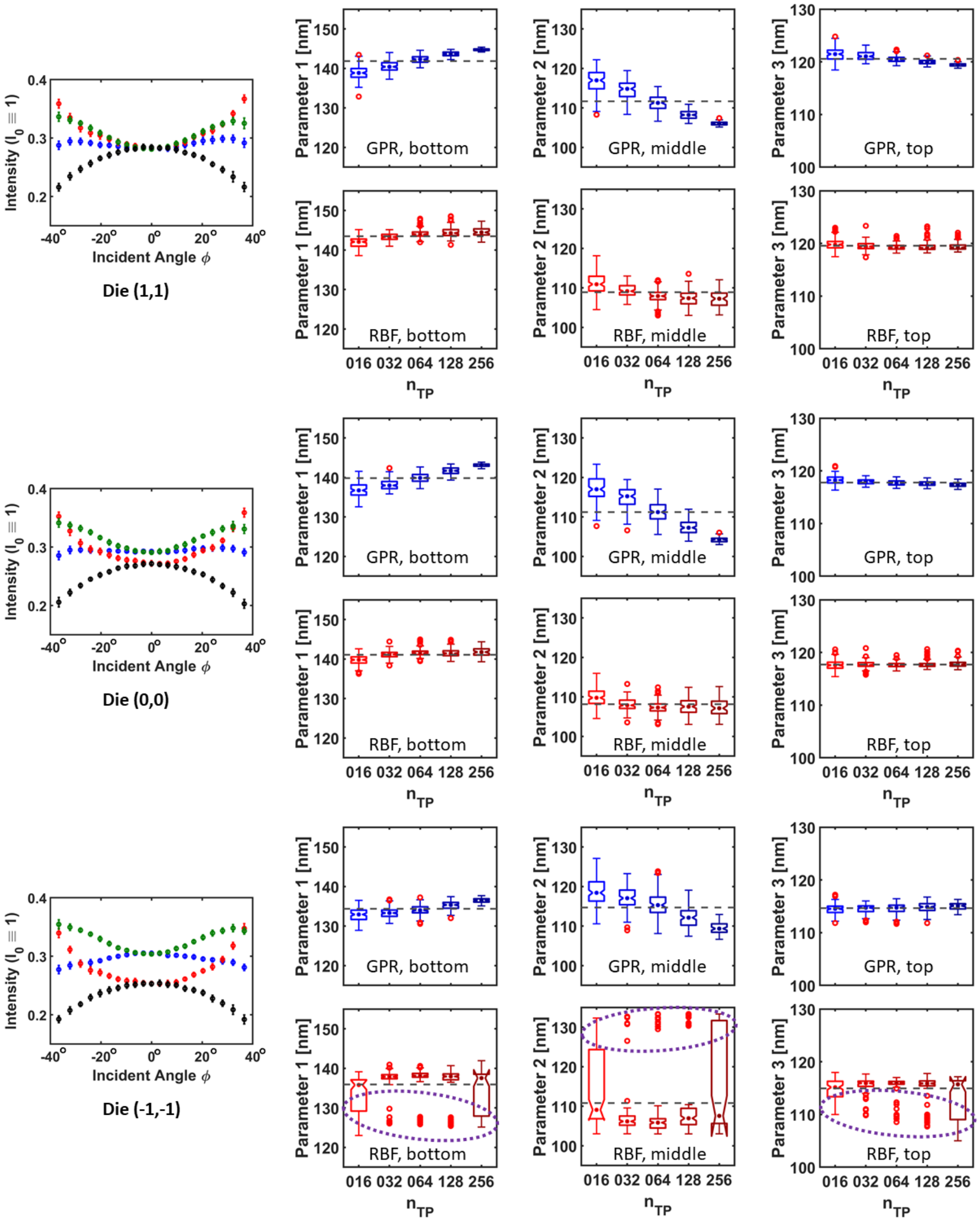


Figure 8. Realizations of the RBF and GPR compared to current library look-up best-fit values. In the left column, measurement error bars are 1σ uncertainties. Box and whisker plots follow definitions in Fig. 4. Ellipses in the bottom-most row indicate the presence of an additional local minimum in the global optimization.

Second, it has been reported that for Bayesian optimization, an objective function can be added to the maximum posterior estimate by means of Bayes' theorem as prior knowledge about non-physical self interactions (e.g., to prevent nonphysical geometries).[17] While the parametric values at this local minimum are physical, they contradict our prior knowledge of these samples from SEM and AFM measurements, and a similar penalty term can be applied. No such function was applied here to illustrate the challenges even for RBF of the inverse problem.

6. CONCLUSION

The need for robust, quantitative machine learning remains especially as sample complexity, improved electromagnetic modeling, and added parameters increase electromagnetic simulation time. While ML can add computational time, this may be smaller in comparison to simulation requirements. Two general approaches have been analyzed, with RBFs augmenting the use of simulated data, while GPR as presented was used to avoid direct application of simulation data. Conventional and ML techniques proved to generally agree on parametric values, but it is clear from the validation steps that uncertainties from GPR here are more problematic than those from RBFs. Testing on experimental data yielded general agreement but a local minimum complicated the fitting in RBF. There are potential solutions to be found for decreasing GPR uncertainty and for constraining the range of allowed fits in RBF due to prior information. These both suggest that additional study is warranted.

Acknowledgements

The authors thank Rick Silver of NIST for supervising the collection of these data and Hui Zhou, formerly of NIST, for initial data processing of the L100P300 data.

REFERENCES

1. den Boef, A. J., "Optical wafer metrology sensors for process-robust cd and overlay control in semiconductor device manufacturing," *Surface Topography: Metrology and Properties* **4**(2), 023001 (2016).
2. Barnes, B. M., Howard, L. P., Jun, J., Lipscomb, P., and Silver, R. M., "Zero-order imaging of device-sized overlay targets using scatterfield microscopy," *Proc SPIE* **6518**, 65180F (2007).
3. Crimmins, T. F., "Defect metrology challenges at the 11nm node and beyond," *Proc SPIE* **7638**, 76380H (2010).
4. Barnes, B. M., Sohn, Y.-J., Goasmat, F., Zhou, H., Silver, R. M., and Arceo, A., "Scatterfield microscopy of 22-nm node patterned defects using visible and duv light," *Proc SPIE* **8324**, 83240F (2012).
5. Silver, R. M., Barnes, B. M., Attota, R., Jun, J., Filliben, J., Soto, J., Stocker, M., Lipscomb, P., Marx, E., Patrick, H. J., et al., "The limits of image-based optical metrology," *Proc SPIE* **6152**, 61520Z (2006).
6. Pomplun, J., Burger, S., Zschiedrich, L., and Schmidt, F., "Adaptive finite element method for simulation of optical nano structures," *physica status solidi (b)* **244**(10), 3419–3434 (2007).
7. Sohn, Y. J., Quintanilha, R., Barnes, B. M., and Silver, R. M., "193 nm angle-resolved scatterfield microscope for semiconductor metrology," *Proc SPIE* **7405**, 74050R (2009).
8. Silver, R., Zhang, N., Barnes, B., Zhou, H., Heckert, A., Dixson, R., Germer, T., and Bunday, B., "Improving optical measurement accuracy using multi-technique nested uncertainties," *Proc SPIE* **7272**, 727202 (2009).
9. Rana, N. and Archie, C., "Hybrid reference metrology exploiting patterning simulation," *Proc SPIE* **7638**, 76380W (2010).
10. Silver, R. M., Zhang, N. F., Barnes, B. M., Zhou, H., Qin, J., and Dixson, R., "Nested uncertainties and hybrid metrology to improve measurement accuracy," *Proc SPIE* **7971**, 797116 (2011).
11. Zhang, N. F., Silver, R. M., Zhou, H., and Barnes, B. M., "Improving optical measurement uncertainty with combined multitool metrology using a Bayesian approach," *Applied Optics* **51**(25), 6196–6206 (2012).
12. Henn, M.-A., Silver, R. M., Villarrubia, J. S., Zhang, N. F., Zhou, H., Barnes, B. M., Ming, B., and Vladár, A. E., "Optimizing hybrid metrology: rigorous implementation of Bayesian and combined regression," *Journal of Micro/Nanolithography, MEMS, and MOEMS* **14**(4), 1 – 8 (2015).

13. Bischoff, J., Bauer, J. J., Haak, U., Hutschenreuther, L., and Truckenbrodt, H., "Optical scatterometry of quarter-micron patterns using neural regression," *Proc SPIE* **3332**, 526–537 (1998).
14. Rana, N., Zhang, Y., Kagalwala, T., and Bailey, T., "Leveraging advanced data analytics, machine learning, and metrology models to enable critical dimension metrology solutions for advanced integrated circuit nodes," *Journal of Micro/Nanolithography, MEMS, and MOEMS* **13**(4), 041415 (2014).
15. Hammerschmidt, M., Weiser, M., Santiago, X. G., Zschiedrich, L., Bodermann, B., and Burger, S., "Quantifying parameter uncertainties in optical scatterometry using Bayesian inversion," *Proc SPIE* **10330**, 1033004 (2017).
16. Heidenreich, S., Gross, H., and Bär, M., "Bayesian approach to determine critical dimensions from scatterometric measurements," *Metrologia* **55**(6), S201 (2018).
17. Hammerschmidt, M., Schneider, P.-I., Santiago, X. G., Zschiedrich, L., Weiser, M., and Burger, S., "Solving inverse problems appearing in design and metrology of diffractive optical elements by using Bayesian optimization," *Proc SPIE* **10694**, 1069407 (2018).
18. Schneider, P.-I., Hammerschmidt, M., Zschiedrich, L., and Burger, S., "Using Gaussian process regression for efficient parameter reconstruction," *Proc SPIE* **10959**, 1095911 (2019).
19. Liu, H., Cai, J., and Ong, Y.-S., "Remarks on multi-output Gaussian process regression," *Knowledge-Based Systems* **144**, 102 – 121 (2018).
20. Henn, M.-A. and Zhang, N.-F., "Model-Based Optical Metrology in R: M.o.R.." `http:/doi.org/10.18434/T4/1502429`. Accessed: 2020-02-28.
21. Sohn, M. Y., Barnes, B. M., and Silver, R. M., "Design of angle-resolved illumination optics using nonimaging bi-telecentricity for 193 nm scatterfield microscopy," *Optik (Stuttgart)* **156**, 635–645 (2018).
22. Silver, R. M., Barnes, B. M., Attota, R., Jun, J., Stocker, M., Marx, E., and Patrick, H. J., "Scatterfield microscopy for extending the limits of image-based optical metrology," *Applied Optics* **46**(20), 4248–4257 (2007).
23. Barnes, B. M., Henn, M.-A., Sohn, M. Y., Zhou, H., and Silver, R. M., "Appraising the extensibility of optics-based metrology for emerging materials," *ECS Transactions* **92**(1), 73 (2019).
24. Yoshioka, S., Matsuhana, B., Tanaka, S., Inouye, Y., Oshima, N., and Kinoshita, S., "Mechanism of variable structural colour in the neon tetra: quantitative evaluation of the venetian blind model," *Journal of the Royal Society Interface* **8**(54), 56–66 (2011).
25. Ehret, G., Pilarski, F., Bergmann, D., Bodermann, B., and Buhr, E., "A new high-aperture 193 nm microscope for the traceable dimensional characterization of micro-and nanostructures," *Measurement Science and Technology* **20**(8), 084010 (2009).
26. "Certain commercial materials are identified in this paper in order to specify the experimental procedure adequately. such identification is not intended to imply recommendation or endorsement by the national institute of standards and technology, nor is it intended to imply that the materials are necessarily the best available for the purpose."
27. Zhang, N. F., Barnes, B. M., Zhou, H., Henn, M.-A., and Silver, R. M., "Combining model-based measurement results of critical dimensions from multiple tools," *Measurement Science and Technology* **28**(6), 065002 (2017).
28. Mongillo, M., "Choosing basis functions and shape parameters for radial basis function methods," *SIAM Undergraduate Research Online* **4**(190-209), 2–6 (2011).
29. Eberhart, R. and Kennedy, J., "A new optimizer using particle swarm theory," in [*Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995. MHS'95.*], 39–43, IEEE (1995).
30. Silver, R. M., Barnes, B. M., Heckert, A., Attota, R., Dixson, R., and Jun, J., "Angle resolved optical metrology," *Proc SPIE* **6922**, 69221M (2008).

Work in Progress: Towards Usable Updates for Smart Home Devices

Julie M. Haney and Susanne M. Furman

National Institute of Standards and Technology (NIST)**, Gaithersburg, MD, USA
{julie.haney, susanne.furman}@nist.gov

Abstract. ***Background.*** Smart home device updates are important tools for remediating security vulnerabilities.
Aim. We aim to understand smart home users' perceptions of and experiences with updates.
Method. We interviewed 40 smart home users and analyzed a subset of data related to updates. We are also planning a broader, follow-on survey.
Results. Users experienced inconsistency in update transparency and methods, were confused about how and if updates are applied, and seldom linked updates to security.
Conclusion. Our efforts will provide a new understanding of smart home updates from a usable security perspective and how those are similar/different to views on updates of conventional IT.

Keywords: Smart home · Updates · Cybersecurity · Usability.

1 Introduction

Internet of things (IoT) smart home updates are a critical mechanism by which manufacturers can distribute patches to remediate security vulnerabilities. Updates may be one of the few tools users have to secure their devices since other configurable security options are limited or unavailable. Unfortunately, technologists have found that update mechanisms may be inconsistent across devices [8]. Even among security professionals, the number one threat to IoT was viewed as "difficulty patching Things, leaving them vulnerable" [16]. Despite technology experts identifying issues, the user perspective on smart home updates has not yet been fully explored.

To better understand experiences and challenges with smart home updates, we analyzed a subset of data from a broader, in-depth interview study of 40 smart home users aimed at investigating general experiences with, perceptions of, and opinions about smart home devices, including aspects of privacy and security. This paper focuses on analysis of update-related data only. By exploring this subset of the interview data, we begin to gain insights into perceptions and

** Certain commercial companies/products are identified in this paper to foster understanding. This does not imply recommendation or endorsement by NIST.

Haney, Julie; Furman, Susanne M. "Towards Usable Updates for Smart Home Devices." Paper presented at 10th International Workshop on Socio-technical Aspects in Security, Guildford, UK. September 17, 2020 - September 17, 2020.

usability of smart home updates, including what role, if any, users perceive updates as playing with the security of their devices. Preliminary analysis suggests that users experience inconsistency in update transparency and methods, as well as confusion about how and if updates are applied. More concerning, most study participants did not relate smart home device updates to security, so they might not have been as inclined to install updates immediately.

Since updates were not a major focus of the interview study, we wish to delve deeper into user update experiences and perceptions, especially on a per-device basis. To that end, we are planning a follow-up survey to gather responses from a broader population of smart home owners. When completed, we hope our research will have several contributions. We will provide novel insights into end user perceptions, experiences, and challenges with updates within the context of smart home devices from both a usability and security perspective. In addition to identifying similarities to prior research focused on updates of other types of computing devices, we hope to discover ways in which smart home device updates may be different or more challenging. Our results may also inform the design of smart home device update mechanisms and notifications to provide a more usable platform for deploying critical security patches when necessary.

2 Related Work

2.1 User Update Behaviors

While no prior studies have explored update behaviors for smart home technologies, researchers have investigated these behaviors for other information technology (IT). People delay software updates for a number of reasons, including a lack of awareness of the upgrade value; interruption of computing activities; and possible negative consequences of applying the update [6, 18]. Users may also have a difficult time understanding the relationship between software updates and security [6]. Ultimately, users must balance the risk and costs of updating against potential benefits [19].

2.2 IoT Updates

A number of critical security vulnerabilities for smart home devices have been identified in recent years, highlighting the need for timely updates [2]. However, there are unique challenges to IoT updates [9]. IoT manufacturers may be inexperienced with security feature and update mechanism design. Economic incentives for providing updates and long-term support for inexpensive and disposable devices may not exist, leaving devices vulnerable to attack. NIST discovered that information on IoT updates is not always readily available to consumers and that updates are not always done in a secure manner [8]. From a technology perspective, IoT devices are often memory, processor, and battery constrained, making updates more challenging to deploy while managing integrity and confidentiality of the updates and potential software dependencies [1, 11, 12].

Several researchers focused on security labels for IoT products. Emami-Naeini et al. [4] showed consumer openness to IoT privacy and security labels, including update information. Morgner et al. [15] investigated consumer preferences for security update information on mandatory IoT product labels. They concluded that security update labels, especially those focused on the availability period (how long the manufacturer guarantees to provide updates) may have a significant impact on consumer product selection.

Although the technical limitations of IoT updates and potential of labels have been discussed, to the best of our knowledge, no prior literature addresses potential usability issues with *smart home* updates through the eyes of consumers, a gap our study hopes to address. Lin and Bergmann [14] suggested that smart home devices should implement updates with little or no user intervention. Emami-Naeini et al. [4] interviewed smart home users, noting that most desired automatic updates because of convenience. However, they made no further observations for recommendations with respect to updates. Other researchers explored user perceptions of smart home privacy and security but did not discuss updates(e.g., [17, 21, 20].

3 Methodology

From February to June 2019, we interviewed 40 smart home users to understand their perceptions of and experiences with smart home devices. NIST's Research Protections Office approved the study. Prior to the interviews, we informed participants of the study purpose and how data would be protected with generic identifiers (e.g., P14_U) not linked to individuals.

3.1 Participant Recruitment and Demographics

We hired a consumer research company to recruit adult users of smart home devices from a database of individuals living in a large U.S. metropolitan area who had agreed to be contacted about research opportunities. To determine eligibility, prospective participants completed an online screening survey about their smart home devices, their role with the devices (e.g., administrator, user), and other demographic information. After reviewing the screening information, we selected participants if they were active users of at least two different types of smart home devices. In line with current interview compensation rates in our region, participants were given a $75 prepaid card.

Participants had diverse professional backgrounds with only eight in an engineering or IT field. Thirty-two of the 40 participants had installed and administered their devices (indicated with an A after the participant ID), and eight were non-administrative users of the devices (indicated with a U). Fifty-five percent were male and 45% were female. Seventy percent were between the ages of 30 and 49. Participants were highly educated with 45% having a master's degree or above and another 50% with a BS/BA. All but one participant had three or more individual smart home devices, with 38 having three or more different categories of devices.

Haney, Julie; Furman, Susanne M. "Towards Usable Updates for Smart Home Devices." Paper presented at 10th International Workshop on Socio-technical Aspects in Security, Guildford, UK. September 17, 2020 - September 17, 2020.

3.2 Data Collection and Analysis

We developed a semi-structured interview protocol covering several topics: purchase and general use; installation and maintenance (including updates); privacy; security; and safety. In this paper, we focus only on data related to updates. An IoT content expert who had professionally worked on IoT security in addition to having an extensive, custom smart home, reviewed the interview questions to ensure the use of correct terminology and the consideration of appropriate aspects of smart home ownership. We piloted the interview with four smart home owners from our institution (two device administrators and two non-administrators/users) to determine face validity of questions and language. Based on feedback from the content expert, we added questions for potential "do-it-yourself" users who customize smart home software and hardware to their own specifications (e.g., via writing custom code). After the pilots, minor adjustments were made to to simplify the wording of several questions. Because modifications were minor, the pilot interviews were included in our analyzed data set. After the protocol was finalized, we collected data via 36 additional semi-structured interviews (40 interviews total including pilots) lasting on average 41 minutes. Interviews were audio recorded and transcribed.

We analyzed the interview data using both deductive and inductive coding practices. Initially, each member of the research team individually coded a subset of four interview transcripts using an *a priori* code list based on research questions and open coded for additional concepts as needed. We then met to discuss codes and develop a codebook. Coding then continued until all transcripts were coded by two researchers, who then met to examine and resolve differences in code application and identify relationships and central themes.

4 Methodology

From February to June 2019, we interviewed 40 smart home users to understand their perceptions of and experiences with smart home devices. NIST's Research Protections Office approved the study. Prior to the interviews, we informed participants of the study purpose and how data would be protected with generic identifiers (e.g., P14_U) not linked to individuals.

4.1 Participant Recruitment and Demographics

We hired a consumer research company to recruit adult users of smart home devices from a database of individuals living in a large U.S. metropolitan area who had agreed to be contacted about research opportunities. To determine eligibility, prospective participants completed an online screening survey about their smart home devices, their role with the devices (e.g., administrator, user), and other demographic information. After reviewing the screening information, we selected participants if they were active users of at least two different types of smart home devices. In line with current interview compensation rates in our region, participants were given a $75 prepaid card.

Participants had diverse professional backgrounds with only eight in an engineering or IT field. Thirty-two of the 40 participants had installed and administered their devices (indicated with an A after the participant ID), and eight were non-administrative users of the devices (indicated with a U). Fifty-five percent were male and 45% were female. Seventy percent were between the ages of 30 and 49. Participants were highly educated with 45% having a master's degree or above and another 50% with a BS/BA. All but one participant had three or more individual smart home devices, with 38 having three or more different categories of devices. Appendix B has more detailed demographics along with the types of devices owned by each participant.

4.2 Data Collection and Analysis

We developed a semi-structured interview protocol covering several topics: purchase and general use; installation and maintenance (including updates); privacy; security; and safety (Appendix A). In this paper, we focus only on data related to updates. An IoT content expert who had professionally worked on IoT security in addition to having an extensive, custom smart home, reviewed the interview questions to ensure the use of correct terminology and the consideration of appropriate aspects of smart home ownership. We piloted the interview with four smart home owners from our institution (two device administrators and two non-administrators/users) to determine face validity of questions and language. Based on feedback from the content expert, we added questions for potential "do-it-yourself" users who customize smart home software and hardware to their own specifications (e.g., via writing custom code). After the pilots, minor adjustments were made to to simplify the wording of several questions. Because modifications were minor, the pilot interviews were included in our analyzed data set. After the protocol was finalized, we collected data via 36 additional semi-structured interviews (40 interviews total including pilots) lasting on average 41 minutes. Interviews were audio recorded and transcribed.

We analyzed the interview data using both deductive and inductive coding practices. Initially, each member of the research team individually coded a subset of four interview transcripts using an *a priori* code list based on research questions and open coded for additional concepts as needed. We then met to discuss codes and develop a codebook. Coding then continued until all transcripts were coded by two researchers, who then met to examine and resolve differences in code application and identify relationships and central themes.

5 Preliminary Results

5.1 Update Modes and Notifications

The interviews revealed that update modes may vary from smart home device to device, with some updating automatically and others requiring users to manually initiate updates. In addition, participants discovered available updates in different ways depending on the device. A participant who owned multiple devices said: *"Some of them notify me, others update automatically, and others I'll find out about either through an email or just because I'm kind of monitoring technology news in general"* (P15_A). Another commented:

> *"Some devices will send me a text message... saying that we're going to be updating a device at this time, and it will apply the updates automatically. Other devices, I need to go into their own specialty apps and check what firmware is running and then check for an update. Some devices, I actually have to go to a website and download something, and then my phone, for instance, will update the device"* (P11_A).

Smart home devices that notify users of available updates do so in a variety of ways. Notifications "pushed" to the device's user interface or via the companion app before or after update installation are most common. For example, an owner of a smart doorbell explained how she finds out about updates: *"I see an alert. It says, 'Your Ring doorbell has a new update. Do you want to allow it? Do you want to accept it?' "* (P36_A). Several participants received emails alerting them of available or just-installed updates. Some devices with screen interfaces, such as smart thermostats and televisions, displayed the update notification directly on the device itself. Other smart home owners did not receive push notifications to tell them updates were available. Rather, they had to manually open the companion app and check.

5.2 Update Purpose and Urgency

Participants most often viewed updates as fixing or adding non-security functionality. For example, one participant stated, *"I accept all updates because I believe they'll make things more functional, add new features that I didn't have before"* (P36_A). Interestingly, this perception led to mixed feelings regarding the urgency of applying updates. Several participants who had experienced issues with their devices believed updates were a high priority. A participant who owns a smart video doorbell and security cameras noted that smart home devices *"would have the highest priorities than any of the other apps on my phone... because that's the security of my home"* (P31_A). Another participant talked about experiencing frequent glitches with his devices. Therefore, he viewed regular updates to his devices as being critical:

> *"To me it's not a choice for, at least, internet of things. Sometimes for my computer, I don't update as soon as they tell me I should. I wait for a while to see if anybody reports bad bugs with the new update. I feel that I have to [for a smart home device] in order for it to work at its best"* (P13_A).

However, others thought updates to functionality were lower priority or unnecessary as long as the device appeared to be working properly. A participant described her indifference with respect to updates, *"I don't think that the end user actually really cares. As long as the thing works, it works"* (P40_U). Other participants did not feel they could properly assess the criticality of the update because the manufacturer did not reveal the purpose of the update: *"The information on what the update achieves is unclear"* (P31_A).

5.3 Uncertainty about Update Status

Participants reflected that they may not observe update notifications, do not recall setting an option to automatically install updates, or are not sure if there are configurable options for setting update parameters. These inconsistencies may lead to a sense of uncertainty about whether their devices are being updated or even can be updated. One user remarked about his virtual assistant, *"I don't know when it's [virtual assistant] doing its updates. Like ever. They never ask me. They never prompt me"* (P7_A).

Some participants assumed that the lack of notifications meant that updates must be happening automatically. While possibly true with some devices, this assumption might be flawed for other products. A participant lamented, *"They don't notify me when there's an update. I guess I just kind of assume that they happen as they go. You would think that I'd get an email, but I guess I don't. That might be nice" (P23_A).*

Even though users may have an assumption of automatic updates, the uncertainty due to lack of notification leaves some with a sense of discomfort. For example, one participant stated: *"I'm assuming that updates are being done silently in the background. I don't really know, and it sort of gives the impression that you bought this thing and it's not evolving. . . that it's not expanding and getting new updates"* (P24_A).

5.4 Updates to Apps vs. Updates to Devices

In addition to uncertainty about update status, the interviews revealed that participants often conflated updates to smart home device companion app software (typically installed on a smartphone) with updates to device firmware. They did not realize that updates to apps were not necessarily accompanied by device updates and vice-versa. This was evidenced by participants referencing typical smartphone app update indicators when asked how they know smart home device updates are available. For example, a user of an Android-based phone explained, *"I get a notification. It doesn't say specifically which apps need to be updated. It just says 48 apps need to be updated. Then I go into Google Play, and see my apps, and individually determine which ones I want to update"* (P31_A).

5.5 Update Concerns

Even when update availability was visible, participants voiced concerns about updates causing issues or breaking functionality on their smart home devices. For example, one participant voiced frustration with updates to his smart televisions: *"I've had to reset my TVs many times because the software update didn't work or kind of messed things up"* (P10_A). Updates also have the potential to invalidate previous user configuration settings or necessitate new ones: *"as they come out with updates, particularly significant updates that change the interface, for example, that might be cause for me to go back in and redo some of the settings"* (P15_A).

Two participants expressed concerns about a lack of updates should a manufacturer stop supporting a product. One of these commented,

> *"I would hope that over time the companies that support these devices would continue to update their firmware and basically make them more reliable. I think in some cases that's happened, but I think in other cases the devices just get abandoned"* (P11_A).

5.6 Relationship to Security

Although some updates can be a conduit to fix security vulnerabilities in smart home devices, study participants rarely linked updates to security, with only five mentioning updates in the context of security. Most discussed updates in terms of fixing functionality or adding features. When asked what mitigation actions they take to address any security concerns they might have, only three mentioned applying updates or upgrading products.

Interestingly, two participants recognized the importance of applying updates, but were also concerned about potential security-related consequences. One participant liked that updates to his devices could be done via the internet, but at the same time was concerned because *"it means that someone's reaching in... There's some kind of access from the outside"* (P26_A). Another saw potential for updates to weaken security:

> *"I guess one area where I would be worried about would be adding features that may threaten my privacy and security... I would want to know that the update also gave me the capability of disabling or turning off that feature I might be concerned about"* (P15_A).

6 Discussion

6.1 Comparison to Traditional Updates

We note similarities between our results and those from previous research studies in Related Work. Similarities included: a lack of awareness of the importance of applying updates; a lack of information about the update purpose hindering users' ability to weigh risk and cost against potential update benefits; concern about possible negative consequences of applying updates; and concern about surprise new features being added.

Although similarities exist, we identified several differences in user experiences with smart home updates as compared to updates explored in prior studies. We did not find evidence of concerns about interruption, likely because users do not have the same kind of interactive sessions with smart devices as they would on a tablet, phone, or computer. Our findings additionally suggest that, because devices are often controlled with a mobile companion app, some updates may be overlooked since several participants did not understand the difference between a phone update, an app update, and a device update. We also discovered that participants were concerned about manufacturers discontinuing product support (and therefore, no longer issuing updates) due to the dynamic smart home market. As opposed to updates for more-familiar and widely-used operating systems, applications, and hardware (e.g., those from Apple and Microsoft), our participants were often unaware if updates were available, how to configure automatic updates, or how to check update status. Confusion about update mechanisms may be amplified by the number of smart home devices users own, especially if the products are from various manufacturers with different update models and different modes of notification.

We also acknowledge that the update experience for smart home devices may necessarily have to be different than traditional IT updates because of processing/memory constraints and limited interactive interfaces. Therefore, more research is warranted to investigate a suitable, usable update interface that can accommodate device limitations.

6.2 Informing Usable Updates

Study results may inform more usable update interfaces and mechanisms. Although our focus was on home users, improved update usability can also be especially valuable for IoT administrators in organizations who have to maintain large numbers of devices.

Insufficient information about the purpose and benefit of updates may result in users lacking a sense of urgency about applying updates, especially if devices appear to be working fine. Users may also be uncertain about update status and availability. To help users make informed decisions, manufacturers could provide greater transparency of update purpose and importance of applying an update (perhaps via a criticality rating), which is in concert with Vaniea and Rashidi's recommendation for easy-to-find information on updates [19]. As also recommended by other standards and government organizations [9, 5, 3, 7], manufacturers could be more forthcoming about their update model and support so that users are aware of how update availability will be made known, what actions users should take to install updates, what update configuration and notification options (if any) are available, and how manufacturers will handle discontinuation of product support. Some of these update attributes were addressed in prior work on product labels [4, 15] and showed promise in impacting consumer purchase decisions and providing transparency. However, more research needs to be done to determine whether consumers would even read the labels.

In addition to lack of transparency, many of our participants expressed discomfort or frustration with updates and their ability to control them. Providing additional information on updates can help users feel more confident in their update decisions. In addition, manufacturers could provide options for users to configure automated updates (as recommended in [14]) with configurable notifications of success afterwards. Users could be given options to schedule if and when they receive notifications. To mitigate concerns that updates might break the device or result in unwanted features or settings, devices could support a rollback mechanism, as recommended by others [8, 13, 19]. Users may then be more likely to install an update if they have a way out should there be a problem.

Although we identified issues related to lack of transparency, it must be noted that it is currently unclear as to whether or not consumers would actually read any additional information or in what format they would wish to receive the information. In addition, too much information could be overwhelming and result in user frustration or users just ignoring the information. Therefore, future research should be done to account for consumer preferences.

7 Limitations and Planned Future Work

In addition to typical limitations of interview studies (e.g., self-report and social desirability biases), our study results may have limited generalizability. Our sampling frame of mostly well-educated individuals living in a high-income region in the U.S. may not be fully representative of the global smart home user population. However, our participant population does appear to typify early adopters of smart home devices as identified in industry surveys (for example, [10]).

Our interview study was meant to be exploratory with a goal of identifying areas warranting additional investigation. As such, the interview protocol was broad in covering multiple aspects of smart home ownership and did not focus solely on updates. We also did not ask about updates on a per-device basis (just generally), so are not able to determine if there are different perceptions or experiences depending on the type of device and manufacturer and if some devices are doing a better job at updates than others.

In recognition that more research should be done to delve deeper into users' smart home update experiences, we are in the initial planning phase for an online, quantitative survey of a larger, more diverse sample of smart home users. In addition to asking more questions about perceptions of updates (e.g., importance, purpose), we will obtain per-device experiences and explore what kind of options, if any, users would like in order to gain greater insight and control of update mechanisms. We will also investigate users preferences for update-related information, e.g., what kind of information they would like to receive (if any at all) and desired formats and communication mechanisms.

References

1. Bauwens, J., Ruckebusch, P., Giannoulis, S., Moerman, I., Poorter, E.D.: Over-the-air software updates in the internet of things: An overview of key principles. IEEE Communications Magazine **58**(2), 35–41 (2020)
2. Consumer Product Safety Commission: Status report on the Internet of Things (IoT) and consumer product safety. https://www.cpsc.gov/s3fs-public/Status-Report-to-the-Commission-on-the-Internet-of-Things-and-Consumer-Product-Safety.pdf (2019)

3. Department for Digital, Culture, Media and Sport: Code of practice for consumer IoT security. https://www.gov.uk/government/publications/code-of-practice-for-consumer-iot-security (2018)
4. Emami-Naeini, P., Dixon, H., Agarwal, Y., Cranor, L.F.: Exploring how privacy and security factor into IoT device purchase behavior. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM (2019)
5. ETSI: TS 103 645 Cyber security for consumer internet of things. https://www.etsi.org/newsroom/press-releases/1549-2019-02-etsi-releases-first-globally-applicable-standard-for-consumer-iot-security (2019)
6. Fagan, M., Khan, M.M.H., Buck, R.: A study of users' experiences and beliefs about software update messages. Computers in Human Behavior **51**, 504–519 (2015)
7. Fagan, M., Megas, K.N., Scarfone, K., Smith, M.: NISTIR 8259 foundational cybersecurity activities for IoT device manufacturers. https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8259.pdf (2020)
8. Fagan, M., Yang, M., Tan, A., Randolph, L., Scarfone, K.: Draft NISTIR 8267 Security review of consumer home Internet of Things (IoT) products. Tech. rep., National Institute of Standards and Technology (2019)
9. Federal Trade Commission: Internet of things privacy and security in a connected world. https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf (2015)
10. GfK: Future of smart home study global report (2016), https://www.gfk.com
11. Gupta, H., Oorschot, P.C.V.: Onboarding and software update architecture for IoT devices. In: 17th International Conference on Privacy, Security and Trust (PST). pp. 1–11 (2019)
12. Hernández-Ramos, J.L., Baldini, G., Matheu, S.N., Skarmeta, A.: Updating IoT devices: challenges and potential approaches. In: 2020 Global Internet of Things Summit (GIoTS). pp. 1–5. IEEE (2020)
13. IoT Security Foundation: Secure design best practice guides. https://www.iotsecurityfoundation.org/wp-content/uploads/2019/11/Best-Practice-Guides-Release-2.pdf (2019)
14. Lin, H., Bergmann, N.: IoT privacy and security challenges for smart home environments. Information **7**(3), 44 (2016)
15. Morgner, P., Mai, C., Koschate-Fischer, N., Freiling, F., Benenson, Z.: Security update labels: Establishing economic incentives for security patching of IoT consumer products. In: Proceedings of the 2020 IEEE Symposium on Security and Privacy. pp. 429–446. IEEE (2020)
16. SANS Institute: Securing the Internet of Things survey. https://www.sans.org/reading-room/whitepapers/covert/paper/34785 (2014)
17. Tabassum, M., Kosinski, T., Lipford, H.R.: "I don't own the Data": End user perceptions of smart home device data practices and risks. In: Fifteenth Symposium on Usable Privacy and Security (2019)
18. Vaniea, K., Rader, E., Wash, R.: Betrayed by updates: How negative experiences affect future security. In: Proceedings of the 2014 SIGCHI Conference on Human Factors in Computing Systems (CHI 14). pp. 2671–2674 (2014)
19. Vaniea, K., Rashidi, Y.: Tales of software updates: The process of updating software. In: Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems (CHI 16). pp. 3215–3226 (2016)
20. Zeng, E., Mare, S., Roesner, F.: End user security and privacy concerns with smart homes. In: Thirteenth Symposium on Usable Privacy and Security (2017)

21. Zheng, S., Apthorpe, N., Chetty, M., Feamster, N.: User perceptions of smart home IoT privacy. In: Proceedings of the ACM on Human-Computer Interaction (2018)

A Interview Questions

SECTION A: TERMINOLOGY

1. You may have heard the term "internet of things," or IoT for short. Can you talk a little about what you think the internet of things is?
2. You may have heard the term "smart devices." What about devices makes them "smart?"
3. What does it mean to have a smart home?
4. What do you think is the relationship, if any, between the internet of things and smart devices?

SECTION B: PURCHASE & GENERAL USE

[Review list of smart home devices before beginning this section.]

5. Who was involved in the decision to purchase the smart home devices?
6. What are the reasons the smart home devices were purchased?
   - How did you (or a household member) learn about the devices before buying them?
7. What hesitations, if any, did you have about getting the devices prior to purchase?
8. For what purposes do you use your smart home devices?
9. How do you access the devices – remotely with an app, while physically in the home, or both?
   - *If using a virtual assistant:* How do you access your devices using [insert assistant name]?
   - *If using a hub:* Do you use the hub app to access your devices, or do you use an individual app specific to each device?
10. How do others in your household use the smart home devices?
11. What do you like most about the devices? What are the benefits, if any, of having these devices?
12. What do you like least or dislike about the devices?
13. How have your opinions or expectations of the devices changed, if at all, from the time you first used them until now?
14. What concerns, if any, do you have about the devices?
15. In what ways, if any, have you changed your behaviors because of your smart home devices?
16. In what ways, if any, have you become reliant on your smart home devices?
17. What do the other members of your household think about the smart home devices?
18. Have you had visitors to the home who have had to use the smart home devices?
    - *If yes:* How did they use the devices? What did they think?

19. What smart home devices, if any, have you had in the past, but are no longer using?
    - What are the reasons for no longer using this device?
20. What kinds of things would you like to be able to do with your devices, but haven't, don't know how, or are not sure that you can?
21. What devices would you like to get in the future? For what reasons?

SECTION C: INSTALLATION/TROUBLESHOOTING

22. Who installed the smart home devices?
23. Who administers (configures or maintains) the smart home devices?

For Installers:

24. In general, what was your experience with the installation of the devices?
    - What went well?
    - What didn't go as well?
25. Have you ever had to reinstall a device? If so, what were the reasons for the reinstallation?
26. *If have more than one device:* What has been your experience adding additional devices to the home?

For DIYers:

27. In the screening questionnaire you indicated you build your own or create extensions for your smart home devices and platforms. Can you briefly summarize what you've done?

For Administrators:

28. What configuration changes, if any, have you made to the devices since installation?
    - *If participant makes configuration changes:* How often do you make changes?
29. How do you know that updates are available or needed?
30. How are updates done on your device - automatically or do you have to initiate them?
    - *If manual initiation:* How often do you check for updates?
    - How do you decide whether to update or not update?

For Everyone:

31. How do you try to figure out how to do something new with your devices?
    - What sources do you consult or use?
    - *If have a voice assistant:* What has been your experience, if any, adding new skills to your voice assistant?
32. What kinds of problems, if any, have you encountered while using your smart home devices?
    - How did you go about trying to resolve those problems?

SECTION D: PRIVACY

33. What type of information, if any, do you think the devices are collecting?
    - Which of this information, if any, would you consider to be personal?
34. Where do you think the information goes?
35. In what ways, if any, does your device or the device manufacturer provide a means to control or manage what information is collected and how it is shared?
36. What are your concerns, if any, about how information is collected, stored, and used and who can see that information?
    - In what ways, if any, have you acted to minimize or alleviate some of those concerns?
    - What kinds of actions would you like to be able to take to address your concerns, but haven't, don't know how, or are not sure that you can?
37. Who do you think is responsible for protecting the privacy of information collected by your smart home device?

SECTION E: SECURITY

38. What are your concerns, if any, about the security of your devices?
    - In what ways, if any, have you acted to minimize or alleviate some of those concerns?
    - What kinds of actions would you like to be able to take to address your concerns, but haven't, don't know how, or are not sure that you can?
39. What restrictions, if any, are placed on who in your home can use the devices and what they can do?
40. How do you authenticate to or get into any apps associated with the device?
    - What issues or problems, if any, have you experienced with authentication?
41. Does more than one person in your household use an app to access the same device?
    - Does more than one person use the same account and authentication to access the app?
    - What concerns, if any, do you have with multiple people having access to the app?
42. Who do you think is responsible for the security of your smart home devices?

SECTION F: SAFETY

43. In what ways, if any, do you think the devices contribute to safety?
44. In what ways, if any, do you think the devices might pose a safety risk?

SECTION G: CONCLUSION

45. Is there anything else you'd like to add related to anything we've talked about?

B Participant Demographics

ID	**Gen**	**Age**	**Ed**	**Occupation**	**Device**	**Type**			
					Sec	**Ent**	**Env**	**Appl**	**Asst**
P1_A	F	50-59	M	Liaison	X		X		X
P2_A	M	30-39	M	Lead engineer	X	X	X		X
P3_A	F	40-49	M	Professor	X	X	X	X	X
P4_A	M	60+	M	Retired	X	X			
P6_U	F	30-39	B	Events manager	X	X	X	X	X
P7_A	M	30-39	B	Software engineer	X	X	X	X	X
P8_A	M	30-39	B	Federal employee	X	X	X	X	X
P9_A	F	30-39	M	Educationist	X	X	X		X
P10_A	M	30-39	B	Computer scientist	X	X	X	X	X
P11_A	M	50-59	M	Electrical engineer	X	X	X		X
P12_U	F	30-39	M	Administrative assistant	X	X	X		X
P13_A	M	50-59	M	Manager, cognitive scientist	X	X	X	X	X
P14_U	F	40-49	H	Information specialist	X	X	X		X
P15_A	M	30-39	B	Computer scientist	X	X	X		
P16_A	M	40-49	M	Research chief	X	X	X		X
P17_A	F	30-39	M	Systems engineer	X	X	X	X	X
P18_A	M	30-39	B	Business consultant	X	X	X		X
P19_A	M	50-59	B	Retail services specialist	X	X	X	X	X
P20_A	F	30-39	B	Administrator		X			
P21_U	F	18-29	B	Human resources manager	X	X	X	X	X
P22_A	M	30-39	B	Executive admin assistant	X	X	X	X	X
P23_A	F	40-49	M	Community arts specialist	X	X	X		X
P24_A	M	40-49	B	Operational safety analyst		X	X		X
P25_A	M	30-39	B	Program management analyst	X	X	X	X	X
P26_A	M	30-39	B	Analyst	X	X	X		X
P27_A	F	40-49	M	Program coordinator	X	X	X	X	X
P28_A	F	50-59	B	Consultant	X		X		X
P29_A	M	18-29	M	Events coordinator	X	X	X		X
P30_U	F	18-29	B	Event planner	X	X	X		X
P31_A	F	30-39	M	Lobbyist	X	X	X		X
P32_A	M	30-39	B	Health educator		X	X	X	X
P33_A	M	18-29	B	Senior technology analyst	X	X	X		X
P34_A	M	40-49	B	Financial analyst	X	X	X	X	X
P35_A	M	40-49	M	Accountant	X	X	X	X	X
P36_A	F	30-39	B	Project manager	X	X	X		X
P37_A	F	40-49	M	Assistant principal	X	X	X		
P38_U	F	60+	M	Special educator		X	X		X
P39_U	M	60+	M	Retired		X	X		X
P40_U	F	30-39	C	Customer service rep	X	X	X		X
P41_A	M	40-49	B	Security	X	X	X		X
				Total	**35**	**38**	**38**	**15**	**36**

Table 1. Participant Demographics. ID: A - smart home administrators/installers, U - smart home users; Gen (Gender); Ed (Education): M - Master's degree, B - Bachelor's degree, C - some college, H - High school; Device Type: Sec - Home security, Ent - Home entertainment, Env - Home environment, Appl - Smart appliance, Asst - Virtual assistant

Poster: Defining Actionable Rules for Verifying IoT Security

Kayla E Ibrahim*, Suryadipta Majumdar*, Daniel Bastos† and Anoop Singhal‡
*Information Security and Digital Forensics, University at Albany - SUNY, USA, Email: {keibrahim,smajumdar}@albany.edu
†British Telecom Research Lab, UK, Email: daniel.bastos@bt.com
‡Computer Security Division, National Institute of Standards and Technology, USA, Email: anoop.singhal@nist.gov

***Abstract*—The Internet of Things (IoT) is being widely adopted in recent years. Security, however, has lagged behind, as evidenced by the increasing number of attacks that use IoT devices (e.g., an arson that uses a smart oven, burglary via a smart lock). Therefore, the transparency and accountability of those devices very often become questionable. To that end, formally verifying the system state of those devices against desirable security rules might be a promising solution. However, there is a significant gap between the high-level IoT security recommendations (e.g., NISTIR 8228, NISTIR 8259, OWASP IoT Security Guidance, ENISA Good Practices for Security of IoT, and UK Code of Practice for Consumer IoT Security), and the low-level IoT system data (e.g., sensor data, logs, configurations). This poster aims to bridge this gap by designing an automated technique to define actionable security rules based on those recommendations and enable the security verification of IoT systems.**

***Index Terms*—IoT, security rules, verification**

I. Introduction

The wide-spread adoption of IoT devices is evident in recent years (with the projections of 75.44 billion devices worldwide by 2025 [11]). Most of those devices, however, are reported to suffer from various security threats due to their implementation flaws and misconfigurations [1], [9], [14]; which often question the accountability and transparency of those devices [1], [7]. To address this concern, verifying the system states of IoT devices against a set of security rules might be a promising solution.

However, the existing security standards, e.g., National Institute of Standards and Technology Internal Reports (NISTIR 8228 [7] and NISTIR 8259 [8]), Open Web Application Security Project (OWASP) IoT Security Guidance [10], UK Code of Practice for Consumer IoT Security [6], and European Union Agency for Cybersecurity (ENISA) Good Practices for Security of IoT [5] are intended more for high-level guidelines than for verifying IoT security. For instance, the recommendation *"ensure proper authentication mechanisms"* from OWASP [10] needs to be instantiated to actionable rules, such as *"no smart door opening without PIN"*.

The existing security solutions (e.g., [2]–[4], [14]) in IoT provide an ad-hoc list of rules for various security solutions, such as, application monitoring, intrusion detection, and access control. However, none of these works develops a generic approach to automatically define actionable rules for verifying IoT device security.

This work targets to overcome this limitation of the existing works, and designs a framework to automatically define actionable security rules for IoT. To this end, we first investigate the existing IoT security standards and identify their limitations in verifying IoT security. Then, we present the design and high-level steps of our proposed framework. Finally, we conclude the current status of this work in progress.

II. Challenges in Defining Actionable Security Rules

We investigate several IoT security standards (e.g., NISTIR 8259 [8], OWASP IoT Security Guidance [10], UK code of practice [6], and ENISA good practices [5]), and identify the following challenges in defining security rules from those standards, as they are not specifically designed for this purpose.

- The recommendations in those standards are too high-level and do not include any system specific information; therefore, for deriving actionable security rules, it is essential to obtain the in-depth system knowledge, and and interpret those recommendations in the context of that system knowledge.
- To verify those recommendations using formal tools requires significant effort including interpreting high-level recommendations to low-level security rules, and preparing these rules (e.g., identifying their data sources, and converting them into formal languages) for security verification.

This work aims to bridge this gap and outline the actionable security rules for verification.

III. Threat Model

We assume that IoT devices may have implementation flaws, misconfigurations, and vulnerabilities that could potentially be exploited by malicious entities to violate security rules. To conduct the verification process, our work relies on a remote server or a local hub/gateway. The communication between the devices and our verification server is secure, using their supported end-to-end encryption mechanisms, e.g., Transport Layer Security (TLS). The privacy threats involved with the data sharing of IoT devices are beyond the scope of this research and will be handled in future research through a privacy-friendly verification technique.

Ibrahim, Kayla; Majumdar, Suryadipta; Bastos, Daniel; Singhal, Anoop. "Defining Actionable Rules for Verifying IOT Safety and Security."
Paper presented at 41st IEEE Symposium on Security and Privacy 2020, San Francisco, CA, US. May 18, 2020 - May 20, 2020.

IV. Approach Overview

Fig. 1 shows the high-level design of our proposed solution.

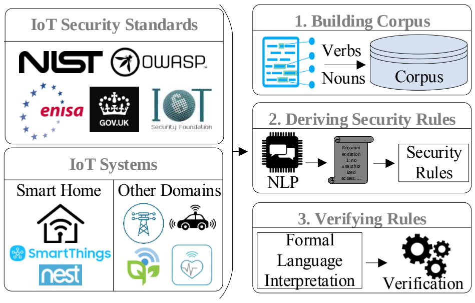


Fig. 1. An overview of our proposed approach

Step 1: Building a Corpus from Security Standards. To build a corpus from the existing IoT security standards, we first parse the contents (i.e., the sections that cover the security guidelines) of those document files. Second, we build a corpus with the relevant terms (i.e., which mainly include the nouns and verbs as those two parts of speech mainly indicate the main message of a recommendation).

Step 2: Deriving Actionable Security Rules. To derive actionable rules, we first extract the key recommendations of those standards by applying several text analytics techniques, such as, term frequency-inverse document frequency (TF-IDF), and natural language processing (NLP) techniques, such as, sentiment analysis [12]. Second, we interpret those key recommendations, apply them in the context of IoT devices, and define actionable security rules for specific cases.

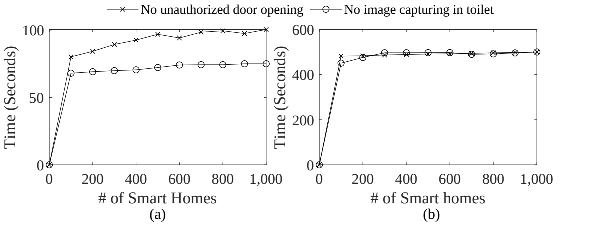
Step 3: Verifying Security Rules. To verify these security rules for actual IoT devices, we translate the actionable security rules into a formal language (e.g., constraint satisfaction problem), collect supporting data for each rule from our smart home testbed, and verify those rules. For verification, we leverage formal verification techniques, e.g., Boolean satisfiability problem (SAT) [13], as it is well-known for its expressiveness, provable security and rigorous results.

V. Preliminary Results

The proposed approach is implemented in a smart home testbed and evaluated for two sample security rules. Fig. 2 shows the total time required for separately verifying the *no unauthorized door opening* and *no image capturing in toilet* security rules. We can easily observe that the execution time is not a linear function of the number of smart homes to be verified. Additionally, our results (not reported here due to space constraint) show that verifying more security rules would not lead to a significant increase in the execution time.

VI. Challenges and Next Steps

While the results of our preliminary experiments indicate the potentiality of leveraging formal tools in IoT security verification, different challenges need to be considered in the next steps of the project. Firstly, the current verification is performed in a remote server, which relies on data sharing and ignores its privacy concerns. Secondly, the Step 2 in Section IV is currently performed manually. Thirdly, there might be domain-specific challenges while adapting our approach in other IoT domains. In the next step, we will explore the feasibility of conducting (fully or partially) the local verification in a hub or gateway; which may require simplifying the workload by developing an incremental approach. Also, we will investigate existing NLP techniques and build an automated technique for Step 2. Finally, we will explore the challenges in applying our approach in other IoT domains.



(a)

(b)

Fig. 2. Total time required to verify two sample security rules, by varying the number of smart appliances to (a) five and (b) 15 in each home.

Disclaimer. This paper is not subject to copyright in the United States. Commercial products are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

References

[1] Omar Alrawi, Chaz Lever, Manos Antonakakis, and Fabian Monrose. SoK: Security evaluation of home-based IoT deployments. In *IEEE S&P*, 2019.

[2] Simon Birnbach, Simon Eberz, and Ivan Martinovic. Peeves: Physical event verification in smart homes. In *ACM CCS*, 2019.

[3] Z Berkay Celik, Patrick McDaniel, Gang Tan, Leonardo Babun, and A Selcuk Uluagac. Verifying internet of things safety and security in physical spaces. *IEEE Security & Privacy*, 17(5):30–37, 2019.

[4] Z Berkay Celik, Gang Tan, and Patrick D McDaniel. IoTGuard: Dynamic enforcement of security and safety policy in commodity IoT. In *NDSS*, 2019.

[5] ENISA. Good Practices for Security of IoT - Secure Software Development Lifecycle, 2019. Available at: https://www.enisa.europa.eu/publications/good-practices-for-security-of-iot-1.

[6] The UK Government. Code of Practice for consumer IoT security, 2019. Available at: https://www.gov.uk/government/publications/code-of-practice-for-consumer-iot-security.

[7] NIST. Considerations for managing internet of things (IoT) cybersecurity and privacy risks, 2019. Available at: https://csrc.nist.gov/publications/detail/nistir/8228/final.

[8] NIST. Recommendations for IoT device manufacturers: Foundational activities and core device cybersecurity capability baseline, 2020. Available at: https://csrc.nist.gov/publications/detail/nistir/8259/draft.

[9] Sukhvir Notra, Muhammad Siddiqi, Hassan Habibi Gharakheili, Vijay Sivaraman, and Roksana Boreli. An experimental study of security and privacy risks with emerging household appliances. In *IEEE CNS*, 2014.

[10] Open Web Application Security Project (OWASP). IoT security guidance, 2019. Available at: https://www.owasp.org/index.php/IoT_Security_Guidance.

[11] Statista. Smart home- United States, Statista market forecast, 2019.

[12] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2), 2011.

[13] Naoyuki Tamura and Mutsunori Banbara. Sugar: A CSP to SAT translator based on order encoding. In *Proceedings of the Second International CSP Solver Competition*, 2008.

[14] Qi Wang, Wajih Ul Hassan, Adam Bates, and Carl Gunter. Fear and logging in the internet of things. In *NDSS*, 2018.

Ibrahim, Kayla; Majumdar, Suryadipta; Bastos, Daniel; Singhal, Anoop. "Defining Actionable Rules for Verifying IOT Safety and Security."

Smart Manufacturing Testbed for the Advancement of Wireless Adoption in the Factory

Richard Candell[1][0000−0002−6679−8823], Yongkang Liu[1][0000−0002−0011−322X], Mohamed Kashef[1][0000−0002−6619−3509], Karl Montgomery[1][0000−0001−5405−6665], and Sebti Foufou[2][0000−0002−3555−9125]

[1] National Institute of Standards and Technology, Gaithersburg, MD
richard.candell@nist.gov http://www.nist.gov
[2] Universite de Bourgogne, France sfoufou@u-bourgogne.fr
www.u-bourgogne.fr/le2i/sebti.foufou

Abstract. Wireless communication is a key enabling technology central to the advancement of the goals of the Industry 4.0 smart manufacturing concept. Researchers at the National Institute of Standards and Technology are constructing a testbed to aid in the adoption of wireless technology within the factory workcell and other harsh industrial radio environments. In this paper the authors present a new industrial wireless testbed design that motivates academic research and is relevant to the needs of industry. The testbed is designed to serve as both a demonstration and research platform for the wireless workcell. The work leverages lessons learned from past testbed incarnations that included a dual robot machine tending scenario and a force-torque seeking robot arm apparatus. This version of the testbed includes computational and communication elements such that the operation of the physical system is noticeably degraded under the influence of radio interference, competing network traffic, and radio propagation effects applied within the lab. The testbed includes two collaborative grade robot arms, programmable logical controllers, and high-performance computing devices for situational tracking, alerting, and control. The paper is aimed to provide our contribution to the exploration of industrial wireless testbed design while soliciting feedback from fellow researchers.

Keywords: factory communications · wireless · testbed design · smart manufacturing · simulation · IIoT.

1 Introduction

One method for measuring the physical effects of communication loss and information delay within an industrial wireless network (IWN) on manufacturing processes within a workcell is the creation of a testbed for conducting research and demonstrating ideas and technologies using real equipment, control systems, and processes. Testbeds are useful for investigating cyberphysical systems (CPS)

Candell, Rick; Liu, Yongkang; Hany, Mohamed; Montgomery, Karl; Foufou, Sebti. "Smart Manufacturing Testbed for the Advancement of Wireless Adoption in the Factory." Paper presented at IFIP 17th International Conference on Product Lifecycle, Rapperswil, CH. July 05, 2020 - July 08, 2020.

impacts in that physical systems are entirely real or at least partially real. A CPS is a tightly coupled integration of physical, computing, and communications processes. As such, events within the computing and communications processes impact the physical process. Within a testbed, physical processes such as robot dynamics do not have to be simulated because gravity, inertia, and system dynamics are inherently present and measurable. Additionally, within a testbed, the electromagnetic properties are present, and their impacts on the physical networking devices are felt and equally measurable in terms of information loss and delay. These losses and delays influence the performance of the network and lead to abnormalities in the performance of the physical system.

CPS Theoretical Model Consulting the theoretical model of a CPS shown in Fig. 1, the physical process instrumented with sensors produces one or more sensor outputs, Y, which are transmitted through a network such as an industrial wireless network. Interference and multi-path phenomena may cause losses and delay as the received sensor information $\bar{Y}$ arrives at the control system. The control system then uses $\bar{Y}$ to act upon the physical system by sending commands, U which may arrive as delayed or lost, demoted by $\bar{U}$, at the physical process. The differences between $\bar{Y}$ and Y and between $\bar{U}$ and U cause changes in the behavior of the physical process. The severity of those changes will depend on the resilience of the control system and the properties of the physical system, itself.

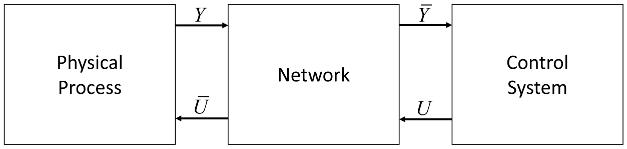


Fig. 1. Theoretical representation of a cyberphysical system.

Motivation While theoretical modeling and simulation are helpful in understanding CPS behaviour, they lack the accuracy of the real-world systems. Therefore, testbeds are used to reproduce larger physical systems on a smaller scale using computing and communication devices and real physical components such as edge devices and robots. The National Institute of Standards and Technology (NIST) has produced a series of industrial wireless testbeds with use cases ranging from chemical production and confined space sensing to robotic machine tending and force controlled actuation. All of these incarnations of the testbed addressed real needs with use cases of low-rate time considerations on the order of 100 ms or more. However, with the emerging requirement of robotic mobility and remote joint actuation in semi-autonomous robotic systems, a new use case is necessary to conduct future research within the testbed. This use case must be selected such that perturbations within the IWN are rapidly reflected within the physical system. The testbed itself should be implemented to measure in real-time both informational and operational performance commensurate with the use case.

Contributions The contributions of this work are as follows: 1) We present our lessons learned conducting experimental studies with past incarnations of the NIST industrial wireless testbed; 2) we present our use case requirements and testbed considerations for a general industrial wireless testbed that readily demonstrates cyberphysical effects rapidly in humanly detectable fashion when information loss and delay are introduced; 3) we present the functional design of the testbed which implements our proposed use case.

Paper Organization The paper is organized as follows: In Section 2, previous relevant academic works are presented. Then in Section 3, lessons learned from past incarnations of software-based and hardware-in-the-loop testbed simulations are reviewed. Next, in Section 4, requirements for the new use case motivated primarily by robotic mobility is addressed along with the requirements for the testbed platform itself. In Sections 5 and 6, the proposed use case and testbed are elaborated in detail. Concluding remarks are given in the final section.

2 Related Work

Wireless communication technologies have been proactively extending their footprints onto the factory floor from non-critical process monitoring services to more deterministic transmission of mission-critical application data [6]. Standardization bodies have released their own visions of service requirements on industrial wireless communications [5, 1]. The European Telecommunications Standards Institute (ETSI) has classified a series of use cases in industrial sectors, mainly as machine-typed communications, and identified corresponding radio transmission requirements on individual classes [5]. The Third Generation Partnership Project (3GPP) has also identied service requirements for industrial wireless use cases and address them in their 5G cellular use scenarios, e.g., massive machine-type communications (mMTC) and ultrareliable and low-latency communications (URLLC) [1]. However, justification of specific values regarding key performance metrics is barely disclosed in these efforts. Recently, there is an ongoing discussion on calibrating wireless service requirements to truly reflect system needs in typical industrial communications scenarios, e.g., factory workcells [12].

Validating industrial wireless requirements and further evaluating candidate solutions have proven challenging. Existing validation approaches can be classified into three main categories: theoretical analysis, software simulation, and hardware testbed. Theoretical analysis has been widely used to showcase performance trend by statistically modelling traffic load and link quality [14]. However, it is very difficult to model complex industrial use cases where factors for both network performance and manufacturing productivity need to be captured thoroughly, especially when transmissions are closely associated with process precision and plant safety. Being widely used to verify theoretical analysis and project system performance, simulation techniques have the flexibility of scaling up and finely tuning the studied use case under different conditions. However, experts in their own domains (e.g., industrial engineers and communications engineers) often find it difficult to select/develop appropriate simulation modules for components that fall out of their expertise, who tend to make over-simplified or erroneous modeling assumptions that deviates the delivery of trustworthy results. Recently co-simulation tools have been developed to jointly mimic behaviors of component systems/networks of CPS in a hierarchical High-Level Architecture (HLA) [13]. Jeng *et al.* introduced a factory co-simulator that jointly considers the production pipeline across multiple workstations and information interaction between them through the network [9]. Such work can also be further extended to the hardware-in-the-loop (HIL) experiments to render more accurate estimation on the performance with tested equipment [3]. Being built upon mathematical models, simulation methods also encounter similar design issues to keep a balance between depiction accuracy and computation efficiency. Besides software-based evaluation, testbeds have been developed to test existing wireless techniques and prototypes in the emulated industrial use cases. Lucas-Estan *et al.* designed experiments to test wireless link redundancy schemes in a multi-robot testbed which is focused on improving reliability of wireless transmissions [11]. Given vast diversity and variability of industrial use cases, the design of effective testbed for individual use cases becomes a key to the success of wireless deployment. NIST engineers have examined the impact of wireless links on industrial system performance through a series of testbeds [3, 7, 8, 4, 2]. Based on lessons learned in these practices, we will examine design principles to highlight industrial wireless requirements in this paper.

3 Lessons Learned from Past Testbed Incarnations

In this section, we discuss the lessons learned from various previous tesbeds that have been used in the NIST industrial wireless systems (IWS) lab. These previously developed testbeds and experimental studies can be found in [3, 7, 8, 4, 2].

3.1 Tennessee Eastman Chemical Factory

In [3], industrial wireless communications devices were deployed for pressure, temperature, and flow sensors data transfer for the control of a simulated chemical factory. IEEE802.15.4-based protocols were deployed such as WirelessHART and ISA100.11a. Both simulations and HIL experiments were performed for this application in [10] and [3], respectively. We found that, generally, process control applications with large time constant can be impacted by communications impairments when the wireless channel is poor for an extended period of time.

The existing WirelessHART and ISA100.11a can perform adequately for these applications while the wireless sensors are connected to the network. However, we found that networks were not robust to events outside the scheme of information transmission such as the loss of a router node, gateway, etc. due to power loss or similar incidents.

Moreover, comparing simulations to HIL experiments, we found that experimental studies are needed to reflect the more realistic impacts of protocol implementation and higher layer protocol impacts such as by Modbus. We have deployed a channel emulator to replicate the impact of electromagnetic propagation in industrial environments on CPS performance. We have also learned that abstract network performance criteria such as packet loss and packet latency may not reflect the communications impact on the production process and hence physical process related criteria should be evaluated in various experimental studies.

3.2 Confined Space Gas Sensing Testbed

In [7], we performed an experimental study on a testbed which is very similar to the chemical factory use case in Section 3.1 such that a similar HIL concept is deployed. The main difference is that different wireless channel models were deployed and different types of information were transferred. In the confined space case, the simulated gas levels at different points of the confined space are transferred over wireless using HIL simulation, and the effectiveness of a safety alerting application is considered. In this setup, we found that the communications protocol implementation could impact the system performance where the implemented Modbus protocol has an update period of 2 seconds which is much slower than the signal and channel variations. We also found that testing can be performed on the signal-level for sensors and actuators signals in case of unavailability of network-level metrics. However, we concluded that the network-level metrics are needed to further understand the reasons behind the measured performance.

3.3 3D Gantry System Testbed

In [8], a commercial 3D gantry system was deployed where the g-code commands to control the gantry system were streamed using a WiFi network. In this experimental study, we found that while converting an originally wired testbed to a wireless one, a ground truth measuring technique has to exist in order to quantify the system performance and hence the wireless impact on performance. Generally, ground truth measurements can be obtained through instrumented feedback from the system, if exists, or through a remote observer where a vision tracking system was deployed in this testbed.

Similar to previously discussed testbeds, the definitions of physical process metrics are needed where these metrics have to be impacted by wireless communications impairments. In this experiment, where g-code command delays made the 3D gantry to halt at a specific location waiting for the following command, performance metrics, such as dwell time and production rate, were considered. The main lesson learned in this experimental study was that queuing of the control commands can mitigate the impact of rapid changes in the wireless channels i.e., keeping a buffered set of commands at the remote station can help to keep the process running. Buffering is not always a viable solution in latency-sensitive applications.

3.4 Force Seeking Apparatus Testbed

In [4], a force seeking apparatus was built using a robot arm and an attached force-torque sensor (FTS). A single wireless channel is deployed where the FTS transmits the force value to the robot controller which reacts when a specific threshold is reached. A machine learning algorithm is used to estimate the interference level on the wireless channel based on tracking the robot arm movements. In this study, we assessed the impact of interference on the physical performance of a fast process where the direct relation between wireless channel quality and the process is quantified. In single wireless channel scenarios, we study only the impact of external interference while competing traffic scenarios are not considered. Hence, we concluded it would be desirable to have a testbed with multiple deployed wireless nodes allowing us to study channel contention scenarios. We also found that the FTSs are not designed to be used wirelessly where large delays can cause dropping the connection with the robot controller. The dropped connection requires resetting the FTS control box to continue its operation which is not acceptable in industrial processes. We concluded that, in general, the industrial devices should be designed to handle information loss and delay more readily through means of mitigative control system programming or user adjustment of device parameters rather than hard-coding total system failure.

3.5 Dual Robot Machine-Tending Testbed

In [2], a more comprehensive testbed was built where two robot arms, one supervisory PLC, and four machine emulators worked together to perform a pick-and-place task. The supervisory control signals to the robot controllers and the machines were transmitted wirelessly. The data for the network traffic and the status of the machines and robot arms were collected. Although, we were able to show the impact of wireless on the network-level metrics, such as packet latency, the impact of wireless on the physical process was almost negligible because only supervisory commands were transferred wirelessly. The latency of supervisory commands were significantly lower than the scan time of the control loop causing negligible physical impact. We discovered that in-network contention is more impactful than electromagnetic interference. Moreover, we discovered that multiple source of contention impacted the system more compared to a single traffic source.

We have also run experiments using both periodic and event-based Modbus polling for the communications between the supervisor and the robot controllers. We found that periodic Modbus polling is more robust to the channel because no timeout warnings may occur. However, it is less bandwidth efficient than the event-based counterpart. On the other hand, event-based Modbus polling is preferred in wireless communications scenarios not to interfere with other network activities. Moreover, access to the robot arm middleware would allow more control of Modbus polling properties. As a result, a middleware alternative such as robot operating system (ROS) would be suitable.

4 Requirements for New Use Case

To enable adoption of an effective use case for deployment in the NIST industrial wireless testbed, development of clear requirements is a necessary and an essential first step. Using our lessons learned and experience, it was required that the new use case proposed have noticeable physical impact under degraded communications situations, have industry relevance, and have an accessible software middleware.

Noticeable Physical Impact Learned from Section 3.5, the new use case requires noticeable physical impact from the degradation of communication. Such degradation occurs commonly in wireless communications. In the use case, wireless links are used to carry traffic to control the physical action directly; thus, any significant fluctuation in latency or reliability of the wireless link will be immediately noticeable by measurement or observation. In past incarnations, physical manifestations from communication loss and delay were not immediately noticeable because the physical systems were more immune to information disturbances in the wireless network.

Industry Relevance Constructing a use case that is similar to the types of applications used by industry is required as the findings discovered will be relevant. This requirement suggests that a new use case should perform a similar task or physical operation that is found in industrial applications.

Access to Software Middleware Many current industrial devices do not support wireless communications in their design. Such devices have proven to have difficulties in compatibility with current wireless solutions, e.g., overly strict timeout thresholds, such as with Modbus client configurations, which may trigger system failures as described in Sections 3.4 and 3.5. Therefore, it is asserted that the desired industrial wireless testbed provide access to the software residing between applications and the industrial wireless networking stack on each host.

5 Proposed Use Case

In this section, we explain briefly the proposed dual robot leader-follower use case. We first introduce the design of the use case where the major system components are defined. Then, we validate that the proposed use case satisfies the use case requirements in Section 4.

5.1 Design

In Fig. 2, the physical process is shown where two robot arms equipped with FTSs are collaborating in lifting a semi-rigid object. Dual robot object lifting is a general industrial approach used for heavy lifting and over-sized material handling. Typically, stationary robots used for lifting communicates through wired technologies. However, the need for flexible manufacturing and mobile robots in future factories requires the deployment of wireless technologies for communications between robots. In the leader-follower schemes, one robot is assigned the rule of the leader to achieve a certain control objective. The other robot, the follower, takes specific actions relative to the leader to maintain a desired relation with the leader with respect to the position or forces on the lifted object. As a result, communications between the leader and the follower plays a crucial role in maintaining the integrity of the lifted object and preventing collisions for the mobile robots.

The control of the leader-follower scenario allows the follower to take decisions based on transferred information about the leader activity. The leader information is transferred using a deployed industrial wireless network. Furthermore, the control of the leader-follower scenario could be done in a centralized approach. In the centralized approach, the centralized controller makes decisions about the follower actions based on the joint state information transferred by the both robots. Moreover, we attach an FTS to the lifted object in order to measure the applied forces and study the impact of wireless communications on the lifted object.

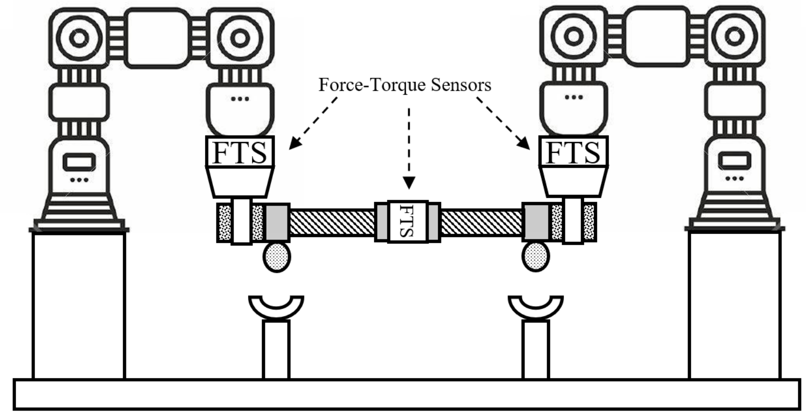


Fig. 2. The leader-follower use case diagram.

5.2 Requirement Coverage

This proposed use case was selected by incorporating the lessons learned from previous use cases in Section 3 and to satisfy the discussed use case requirements in Section 4. In the proposed use case, the goal is to get a direct impact of wireless communications on the physical process through having the leader-follower relation between the robots. Any latency or reliability issues in the information transfer between the leader and the follower is expected to affect the desired relation between them. The impact of wireless communications on the physical process can be measured through the position feedback of the robot arms and the FTSs. Capture of network data is made possible through the use of network probes strategically placed throughout the testbed. Similarly, capture of performance data of the physical system is made possible by instrumenting the testbed with ground truth measurement and recording probes such that variables such as forces, torques, positions, velocities, and accelerations are captured independently of the wireless network. We deploy a master clock to synchronize all the collected physical process and network traffic data. Finally, this proposed use case replicates a realistic scenario which can be found in heavy lifting and over-sized material handling.

6 Proposed Testbed

6.1 Architecture

The testbed serves as an evaluation platform that emulates workcell operations in the proposed use case and measures the impact of node interactions on the production performance. As shown in Fig. 3, it consists of three main subsystems according to their different roles: *workcell modules*, *network components*, and *measurement devices*. In the designed architecture, subsystems are connected with each other through predefined interfaces and integrated as a complete evaluation process. Design details of each one will be elaborated in the following section.

6.2 Components

Workcell modules include testbed components that participate in manufacturing tasks, e.g., robots and their controllers in the leader-follower use case. In the testbed, we deploy two Universal Robots (UR) UR3 CB-series robots that serve as actors in the leader-follower use case. Each UR3 is equipped with a 6 degree-of-freedom (6-DoF) robot arm, one control box, and peripheral devices, known as URCaps, which include a Robotiq gripper and a 6-axis force-torque sensor attached to the end effector. Instead of using proprietary UR programs, the control of robots implements agents atop the robot operating system (ROS) middleware which provides a rich set of options in control functions and peripheral support for UR3. In ROS, robot-oriented processes are identified as separate

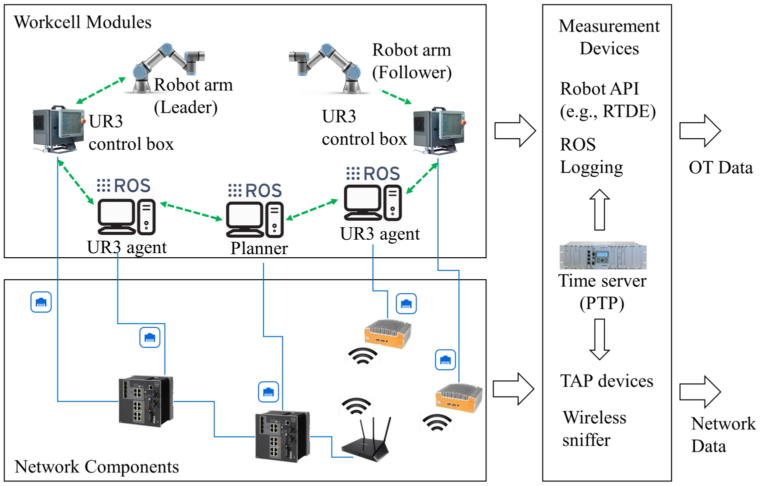


Fig. 3. Workcell testbed architecture

function nodes which can be distributed across multiple physical machines being connected through the network. As shown in Fig. 3, the testbed hosts three computers running Ubuntu 18.04 with the ROS Melodic distribution. Two of them serve as UR3 agents where ROS UR3 drivers are installed to adapt universal ROS functions and commands to UR3 features. The coordination between the leader and the follower can be realized through the peer-to-peer communications between UR3 agents. In this case, the third machine, i.e., the ROS planner, performs motion planning for both robots. There is another feasible approach in which both UR3 agents report to the planner who will resolve their joint trajectory actions and reply to both with respective action commands.

Network components are network devices that enable workcell modules to communicate with each other for job-oriented information exchange and process control. Each workcell module in the testbed contains one or more built-in Ethernet interfaces. Two Cisco IE-4000 industrial gigabit Ethernet switches serve as the wired network backbone. Ethernet-wireless adapters are used to replace wired connections by enabling wireless communications between modules. Fig. 3 illustrates an example of connecting to the follower robot through wireless. The testbed uses IPv4 addresses to manage communication devices. Additional network connections are also planned in the testbed that provide measurement data links and other complimentary features, such as time synchronization services.

Measurement devices collect data including but not limited to manufacturing progress, machine status, link quality, and network health. Table 1 summarizes

Table 1. Testbed measurement data collections

Data	Source	Type	Time Synced	Format
Actual joint status (e.g., position, velocity, and acceleration)	UR3 joints	Workcell	Yes, PTP	CSV
URCap force and torque	FTSs at UR3 end effectors	Workcell	Yes, PTP	CSV
ROS UR3 control (Leader/follower status)	ROS UR3 agents	Workcell	Yes, PTP	CSV
ROS Motion planning output (target joint status)	ROS master	Workcell	Yes, PTP	CSV
Force and torque (ground truth)	FTS at the lifted object	Workcell	Yes, PTP	CSV
Ethernet packet captures	TAP	Network	Yes, PTP	PCAP
Wireless packet captures	Wireleess sniffer	Network	Yes, PTP	PCAP

data to be collected and saved for further analysis. Specifically, production-related workcell data and network data are monitored and captured at different measurement devices. The former ones are usually collected through UR3's API, e.g., the real-time data exchange (RTDE), or saved locally in workcell modules, e.g., the ROS logging at ROS nodes; the latter ones are mainly captured by traffic analysis tools, such as test access point (TAP) devices for real-time traffics in Ethernet links and the wireless sniffer. Note that the testbed uses real-time readings from the FTS at the lifted object as the ground truth for measuring the performance of coordinated operations between the leader and the follower. Besides, timing information is also rendered and embedded into each record so that discrete events can be aligned in time to review the operation logic and network behaviors in future analysis. Therefore, the testbed can implement the precision time protocol (PTP) to provide the sub-microsecond level accuracy across the networked distributed clocks. A Meinberg M900 time server is used as the central PTP master clock that supports the IEEE 1588-2008 (PTPv2) standard.

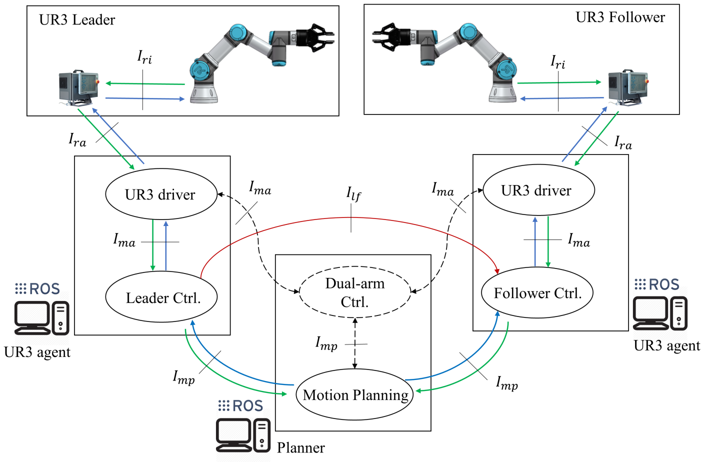


Fig. 4. Testbed interfaces

6.3 Interfaces

Fig. 4 illustrates interfaces between workcell modules that are used to enable the coordination between the two actors in the proposed leader-follower use case. As data communications are directional, we define the *upstream* data flow to be the one from the end node, i.e., the UR3 robot arm, toward the planner which is denoted by green arrows in the figure. The *downstream* data is reversed and denoted in blue. When the leader(follower) has its own controller residing in the agent machine, the leader will routinely share its motion information to the follower, either as geometrical variables or force-torque readings, which allows the follower to track the leader's trajectory and calculate its own route to keep the metrology bar in the desired state. In another case when the ROS planner also takes over the control mission, a new dual-arm control node will replace the separate controllers in UR3 agents. The flows will be updated accordingly which are illustrated as dashed arrows in Fig. 4.

Table 2 summarizes interfaces defined for data exchanges between different nodes. Note that each interface contains data in both upstream and downstream directions. Except I_{ri} for UR3 internal traffics, the other interfaces are built for communications between ROS nodes. ROS provides three communication styles including synchronous request/reply interactions ("services"), asynchronous data streaming ("topics"), and data storage in a parameter server. We will use the testbed to explore the capability of wireless links in support of different communications styles in selected interfaces. For each interface, a set of protocols is listed for which the use case will be tested. Ethernet is used for baseline performance testing, while wireless local area network (WLAN) protocols

Table 2. Testbed interface specifications

Interface	Protocol	Data		Note
		Upstream	Downstream	
UR3 internal, I_{ri}	Serial	UR3 and URCap status	Joint trajectory action	UR3 proprietary
UR3-ROS, I_{ra}	Ethernet WLAN	Robot status	Joint trajectory action	ROS UR3 driver
ROS master-agent, I_{ma}		Robot status	Joint trajectory action	ROS topics/services
ROS motion planning, I_{mp}		Target trajectory	Joint trajectory action	ROS topics/services
Leader-Follower, I_{lf}		Force-torque-based or geometrical trace		ROS topics/services

such as IEEE 802.11 will be used when a wireless protocol is employed. Fifth generation wireless (5G) is envisioned as well when resources come available.

7 Conclusions

In this paper, we have presented the next evolution of an industrial wireless testbed. The testbed is designed with the goal of investigating use cases that will noticeably degrade under the influence of radio interference, competing network traffic, and radio propagation effects applied within the lab. One specific use case is presented in which two robots perform a coordinated lift operation using either a geometric leader follower or a force-torque minimization control scheme. In the control scheme presented, wireless links are used to carry joint state information such that lost or delayed state information immediately impacts physical action. Capture of performance data of the network and physical systems is made possible by the use of probe devices. Our testbed will serve as a research and demonstration platform for reliable industrial wireless deployment and control system strategies. Finally, our work is presented with lessons learned from past incarnations of our wireless testbed such that our work may help others construct better wireless testbeds by learning from our successes and failures. Once completed, results from our testbed will be made available through reports, papers, and publication of data sets.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

1. 3GPP: Service Requirements for the 5G System. Technical Specification (TS) 22.261, 3rd Generation Partnership Project (3GPP) (Dec 2019),

https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx ?specificationId=3107, version 16.10.0

2. Candell, R., Kashef, M., Liu, Y., Montgomery, K., Foufou, S.: A graph database approach to wireless iiot work-cell performance evaluation. In: Proceedings of the 2020 IEEE International Conference on Industrial Technology (Feb 2020)
3. Candell, R., Lee, K.: Measuring the effect of wireless sensor network communications on industrial process performance. In: 2015 ISA process control and safety symposium, Houston, TX (2015)
4. Candell, R., Montgomery, K., Kashef, M., Liu, Y., Foufou, S.: Wireless interference estimation using machine learning in a robotic force-seeking scenario. In: 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE). pp. 1334–1341 (Jun 2019). https://doi.org/10.1109/ISIE.2019.8781418
5. (ETSI): Etsi tr 103 588 reconfigurable radio systems (rrs); feasibility study on temporary spectrum access for local high-quality wireless networks. Tech. rep., European Telecommunications Standards Institute (ETSI), Sophia-Antipolis (2018)
6. Huang, V.K.L., Pang, Z., Chen, C.A., Tsang, K.F.: New trends in the practical deployment of industrial wireless: From noncritical to critical use cases. IEEE Industrial Electronics Magazine **12**(2), 50–58 (Jun 2018). https://doi.org/10.1109/MIE.2018.2825480
7. Kashef, M., Candell, R.: Characterization for an industrial wireless network in a gas sensing scenario. In: Proceedings of the 2017 ISA International Instrumentation Symposium (2017)
8. Kashef, M., Candell, R., Foufou, S.: On the Impact of Wireless Communications on Controlling a Two-Dimensional Gantry System. International Manufacturing Science and Engineering Conference, vol. Volume 1: Additive Manufacturing; Manufacturing Equipment and Systems; Bio and Sustainable Manufacturing (06 2019). https://doi.org/10.1115/MSEC2019-2896, v001T02A005
9. Li, H., Geng, J., Liu, Y., Kashef, M., Candell, R., Bhattacharyya, S.S.: Design space exploration for wireless-integrated factory automation systems. In: 2019 15th IEEE International Workshop on Factory Communication Systems (WFCS). pp. 1–8 (May 2019). https://doi.org/10.1109/WFCS.2019.8757954
10. Liu, Y., Candell, R., Moayeri, N.: Effects of wireless packet loss in industrial process control systems. ISA Transactions **68**, 412–424 (May 2017). https://doi.org/10.1016/j.isatra.2017.02.005
11. Lucas-Estañ, M.d.C., Maestre, J., Coll-Perales, B., Gozalvez, J., Lluvia, I.: An experimental evaluation of redundancy in industrial wireless communications (09 2018). https://doi.org/10.1109/ETFA.2018.8502497
12. Montgomery, K., Candell, R., Liu, Y., Hany, M.: Wireless User Requirements for the Factory Work-cell. Tech. rep. (2019). https://doi.org/10.6028/NIST.AMS.300-8, https://www.nist.gov/publications/wireless-user-requirements-factory-workcell
13. Neema, H., Gohl, J., Lattmann, Z., Sztipanovits, J., Karsai, G., Neema, S., Bapty, T., Batteh, J., Tummescheit, H., Sureshkumar, C.: Model-based integration platform for fmi co-simulation and heterogeneous simulations of cyber-physical systems. pp. 235–245 (03 2014). https://doi.org/10.3384/ecp14096235
14. Zheng, M., Liang, W., Yu, H., Xiao, Y.: Performance analysis of the industrial wireless networks standard: Wia-pa. Mobile Networks and Applications pp. 1–12 (2015). https://doi.org/10.1007/s11036-015-0647-7

Digital Radiometer for Traceable Spectrum Sensing

Xifeng Lu[1, 2], Daniel Kuester[1], and Dazhen Gu[1]
[1]Communication Technology Lab, National Institute of Standards and Technology, Boulder, CO USA 80305
[2]Department of Physics, University of Colorado, Boulder, Colorado 80309, USA
xifeng.lu@nist.gov

***Abstract*—This paper describes a metrology instrument designed for detecting noise and interference in digital communications. A quadrature detector downconverts in-phase and quadrature channels, which are each digitized at high speed. This architecture makes the upper and lower signal sidebands separable by digital processing, and has spectral resolution that is unprecedented in conventional analog detection. Several considerations for design, system characterization and calibrations are discussed.**

***Index Terms*—Measurement, noise, radiometer, sideband separating downconversion, standards.**

I. Introduction

Noise is an ultimate limiting factor in the performance of wireless communication links. With the proliferation of communication devices leading to the worldwide spectrum shortage, the ever-rising noise floor becomes more and more prominent in both licensed and unlicensed frequency bands. An accurate quantification of noise and interference level with its spectral, temporal, and geographic signatures has increasing importance for wireless and radar engineering, spectrum users, and objective assessment by policymakers.

Microwave radiometers are a ubiquitous noise-measurement tool in the metrology community. Conventionally, noise detection has been implemented with analog detectors. Although analog detectors have functioned well in measuring stable noise sources with minimal frequency dispersion, analog detection is less flexible and suffers from limited frequency selectivity, which limits its practical application for modern spectrum users.

At the National Institute of Standards and Technology (NIST), we have started building a metrology-grade radiometer centered on digital detection. By leveraging the existing thermal noise reference standards and rigorous uncertainty analysis, the new instrument will enable us to measure noise and interference signals traceably and with dramatically improved spectral resolution.

II. Design of Digital Radiometer

Although it is possible to directly sample a broadband signal centered at a frequency of interest with a high-speed digitizer, in our new digital radiometer we downconvert from the radio frequency (RF) first to improve tuning flexibility, and energy efficiency, and to reduce cost.

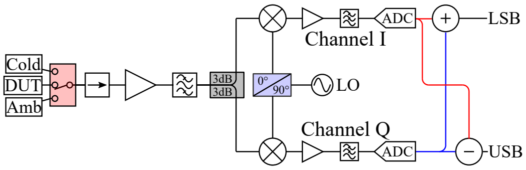


Fig. 1. A block diagram of the digital radiometer with the IQ downconversion configuration.

Conventional radiometer detectors typically use double-sideband (DSB) downconversion. The DSB-downconverted intermediate-frequency (IF) signal is an aggregation of the upper-sideband and lower-sideband spectral components. As a result, the DSB architecture cannot unambiguously distinguish between signals with asymmetrical spectra. In order to preserve the ability to distinguish signals in both sidebands, sideband-separating downconversion needs to be used. In the communication field, this technique is equivalent to in-phase (I) and quadrature (Q) mixing. As shown in Fig. 1, the RF signal to be detected is split into two nearly identical signal-processing chains; the I channel and the Q channel. The only difference between these channels is a 90° offset between the local oscillator (LO) signals driving the IQ mixers, accomplished by the quadrature hybrid divider. The downconverted IQ signals are eventually sampled by a dual-channel analog-to-digital converter (ADC). Next, digital processing is implemented to account for non-idealities of the IQ channels, calibrate the absolute magnitude with the known noise references, and ultimately reproduce the upper- and lower-sideband spectra of the signal.

Before the signal enters the digital domain, a significant amount of amplification is required to bring the low-level thermal noise signal within the ADC dynamic range. The power level of the noise signal associated with the NIST noise reference standards is typically around -90 dBm. More than 100 dB gain may be required to compensate additional conversion and cable losses. Prior to any amplification stages, isolators are used to minimize the variation of the input impedance seen by the radiometer, which is critical in metrology applications. The filters reduce amplifier saturation, suppress out-of-band harmonics and the LO leakage, and eliminate aliasing into the ADC.

III. System Characterization and Calibration

A. Receiver Stability

In light of the significant signal gain in the system, radiometer receiver stability with time is a serious concern. Consequently, frequent calibrations of the digital radiometer

Lu, Xifeng; Gu, Dazhen; Kuester, Dan. "Digital Radiometer for Traceable Spectrum Sensing." Paper presented at Conference on Precision Electromagnetic Measurements (CPEM), Denver, CO, US. August 24, 2020 - August 28, 2020.

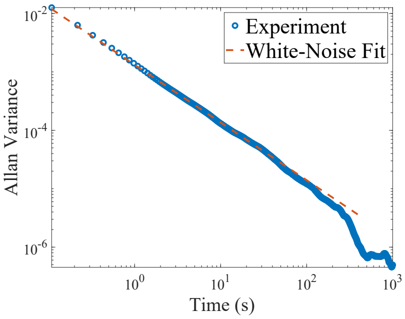


Fig. 2. Allan variance of the digitally sampled data from the radiometer when a thermal noise source was connected.

against known NIST references may be required for a traceable measurement. Measurements of noise reference standards constitute "dead time", during which we cannot measure spectral emissions from a device under test. There is thus a trade-off between accuracy and availability.

To maximize availability, it is important to determine how often the calibration needs to be refreshed. To that end, the Allan variance has been proven a reliable technique to characterize the receiver stability and estimate the calibration intervals. The Allan variance allows characterization of the system drift by identifying frequency-dependent noises other than white noise [1]. Figure 2 shows the Allan plot of the digital-radiometer data and indicates an excellent receiver stability up to more than 50 s in a controlled lab environment. The Allan variance result provided a guidance of the calibration period on the order of about 1 minute.

B. IQ Imbalance

Ideally, the addition of the I and Q signals would produce the lower-sideband spectrum, while the subtraction of the two would produce the upper-sideband spectrum. In practice, the imperfection of analog components leads to differences in the complex gain (a combination of the phase delay and magnitude amplification) between the two channels, which results in IQ imbalance. Because of the IQ imbalance, the simple addition and subtraction of the IQ baseband signals do not cleanly separate of the lower- and upper-sideband spectra. Even meticulous selection of the analog components can only achieve uncalibrated sideband rejection of 20 dB or less [2].

To overcome the IQ imbalance, a calibration can be performed with digital data to improve the sideband rejection. A sine tone at a known frequency can be injected into the digital-radiometer frontend, from which a complex-valued calibration coefficient is chosen to null the digital spectrum output at the image frequency. The coefficients for each of the lower sideband (LSB) and upper sideband (USB) are

$$C(f) = \begin{cases} \frac{X_{\mathrm{I}}(f_{\mathrm{LO}}-f)}{X_{\mathrm{Q}}(f_{\mathrm{LO}}-f)}, & \text{for LSB } f < f_{\mathrm{LO}}, \\ \\ -\frac{X_{\mathrm{I}}(f_{\mathrm{LO}}-f)}{X_{\mathrm{Q}}(f_{\mathrm{LO}}-f)}, & \text{for USB } f > f_{\mathrm{LO}}, \end{cases} \tag{1}$$

where, X_{I} and X_{Q} are the Fourier transformed time-series data in the I and Q channels, respectively. As an example, we can estimate $C(1.8\,\mathrm{GHz})$ from the downconverted IQ baseband signal at -0.3 GHz when the LO frequency is fixed at 1.5 GHz.

C. Spectral Resolution

Application to spectrum sensing simultaneously in multiple frequency allocations requires transformation into the frequency domain. Mathematically, this type of analysis is equivalent to the projection of a time series onto a set of orthogonal basis [3]. The most common method is the discrete Fourier transform (DFT). However, the DFT of a time series that is non-periodic sequence and acquired on a finite interval cause apparent energy leakage in the frequency spectrum of the signal. For a series of size N sampled at a frequency of f_{s}, the DFT spectrum can be expressed by

$$\check{X} = \Pi \circledast X, \tag{2}$$

where $\check{X}$ is the resultant finite-DFT spectrum of the signal, and X is the actual spectrum of the signal. The symbol $\circledast$ denotes convolution, and Π is the Fourier transform of the boxcar function spanning over the observation interval N/f_{s}; i.e. a set of the sinc functions centered at $-\frac{N}{2}\frac{f_{\mathrm{s}}}{N}$, $(-\frac{N}{2}+1)\frac{f_{\mathrm{s}}}{N}, \ldots, (\frac{N}{2}-1)\frac{f_{\mathrm{s}}}{N}$. Spectral components of signals with their base-band representation at other frequencies introduce non-zero projections on all bases, causing the spectral leakage.

To mitigate the spectral leakage problem, a modified basis set is required so that an arbitrary signal would primarily project onto it with its center frequency in the closest vicinity of the signal frequency. This manipulation is equivalent to multiplicatively applying a window to the time series sampled by the ADC. A number of windowing functions will be tailored for applications specific to the communication signals of interest [4]. Window design considerations include the equivalent bandwidth, the flatness within the pass band, the side-lobe peak level. Regardless of what window functions are used, the spectral analysis of the digitally sampled signal will produce much finer spectral resolution than what exists in the analog radiometers.

IV. Conclusion

In summary, we designed a digital radiometer by use of IQ mixing and ADC sampling for measuring noise and interference in communication applications. Taking advantage of existing NIST thermal-noise standards, the digital radiometer enables traceable spectrum measurements. Uncertainty analysis will be shown in the extended paper.

References

[1] W. J. Riley, *Handbook of Frequency Stability Analysis*, NIST Publication 1065, 2008.

[2] R. Finger, P. Mena, N. Reyes, R. Rodriguez, and L. Bronfman, "A calibrated digital sideband separating spectrometer for radio astronomy applications", *Publications of the Astronomical Society of the Pacific*, vol. 125, no. 3, pp. 263–269, March, 2013.

[3] C. W. Helstrom, *Statistical Theory of Signal Detection*, 2nd ed., New York: Pergamon Press, 1968.

[4] F.J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform", *Proceedings of the IEEE*, vol. 66, no.1, pp.51-83, Jan.1978.

The Design of an Instrument to Realize Small Torque at NIST

Leon Chao, Rafael Marangoni, Frank Seifert, Darine Haddad, Jon Pratt, David Newell, and Stephan Schlamminger
National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg MD 20899
leon.chao@nist.gov

***Abstract*—After the recent redefinition of the International System of Units (SI), torque no longer needs to be traceable to a calibrated mass in a gravitational field suspended from a known lever arm and disseminated through a chain of torque transducers.**

An SI-traceable torque can be directly realized using the Kibble principle where a torque is generated by a calibrated electromagnetic transducer. Here, a set of theoretical designs for a new instrument aimed at direct realization of torque to 0.1% accuracy is analyzed and examined for feasibility. With careful attention to magnet design, a robust and easily implemented torque calibrator can be built.

***Index Terms*—Torque metrology, Kibble balance, precision engineering design**

I. Introduction

From aircraft and automobiles down to cell phones and wrist watches, mechanical assemblies are held together with threaded fasteners whose optimal performance and reliability depend on accurate application of torque. For example, six million components of a Boeing 747 are constrained together by millions of threaded fasteners secured using the correct amount of torque, applied by technicians using torque wrenches or torque screwdrivers.

The technical standard ISO 6789 by the International Standard Organization, for example, requires all torque wrenches and screwdrivers to be calibrated every 5000 uses or once per year [1]. These handheld tools are calibrated by torque sensors which typically have relative uncertainties of 0.25% and are traceable to torque transducers ultimately calibrated by deadweight torque machines. Currently, the lowest uncertainty deadweight torque machine in the USA has a relative uncertainty of 2×10^{-5} [2]-[3].

The existing traceability is simple, reliable, and convenient for the most part; however, at torques below 1 N m the handling of small mass artifacts required to calibrate the deadweight machines is difficult, irreproducible, and burdensome.

Torque is expressed as a product of force and length in units of Newton-meter (N m). A deadweight machine relies on a measured force generated by a calibrated mass in a known gravitational field applied at a well-characterized lever arm distance. The recent redefinition of the International System of Units (SI) enables new approaches to traceable realization of units of measure. Mechanical force no longer needs to be traceable to a mass artifact, but rather it can be realized through electromagnetic force, ultimately traceable to quantum standards [5]. Furthermore, by implementing the Kibble principle to be discussed in the next section, the previously required lever arm length measurement becomes obsolete.

The Fundamental Electrical Measurements Group at NIST is investigating the design of a tabletop industrial instrument for realizing torque using the Kibble method, drawing inspiration from [4]. The instrument under consideration aims to realize small torques with a relative uncertainty of 0.1%. The instrument seeks to exploit convenient electrical and time standards traceable to quantum standards to achieve a robust alternative to present cumbersome practices for torques below 1 N m.

II. Theory of the Kibble Principle for rotation

Given a rigid loop of conductive wire with an enclosed area A traveling through a magnetic field with flux density B, the centroid of the loop has velocities v_x, v_y, and v_z and angular velocities about each axis ω_x, ω_y, and ω_z. The total flux Φ through the loop is

$$\Phi = \oiint \vec{B} \cdot d\vec{A}. \tag{1}$$

The induced voltage V, according to Faraday's law of induction [6] that appears across the open terminals of the loop is

$$\begin{aligned} V &= -\frac{\mathrm{d}\Phi}{\mathrm{d}t} \\ &= -\left(\frac{\partial\Phi}{\partial x}v_x + \frac{\partial\Phi}{\partial y}v_y + \frac{\partial\Phi}{\partial z}v_z + \right. \\ &\qquad \left. \frac{\partial\Phi}{\partial\theta_x}\omega_x + \frac{\partial\Phi}{\partial\theta_y}\omega_y + \frac{\partial\Phi}{\partial\theta_z}\omega_z\right). \end{aligned} \tag{2}$$

If assuming the loop is an open circuit and attached to the end of a rotor capable of undergoing rotation $\omega_{z'}$ where z' is aligned to gravity, then each off-axis velocity component is solely determined by $\omega_{z'}$ and can be expressed as

$$\begin{aligned} v_x = \omega_{z'}\frac{\partial x}{\partial\theta_{z'}}, \quad v_y = \omega_{z'}\frac{\partial y}{\partial\theta_{z'}}, \quad v_z = \omega_{z'}\frac{\partial z}{\partial\theta_{z'}}, \\ \omega_x = \omega_{z'}\frac{\partial\theta_x}{\partial\theta_{z'}}, \quad \omega_y = \omega_{z'}\frac{\partial\theta_y}{\partial\theta_{z'}}, \quad \omega_z = \omega_{z'}\frac{\partial\theta_z}{\partial\theta_{z'}}. \end{aligned} \tag{3}$$

Substituting eq. 3 into eq. 2, the following expression for the induced voltage of the loop at angle $\theta_{z'}$ undergoing instantaneous angular velocity $\omega_{z'}$ is obtained

$$\begin{aligned} V &= -\omega_{z'}\left(\frac{\partial\Phi}{\partial x}\frac{\partial x}{\partial\theta_{z'}} + \frac{\partial\Phi}{\partial y}\frac{\partial y}{\partial\theta_{z'}} + \frac{\partial\Phi}{\partial z}\frac{\partial z}{\partial\theta_{z'}}\right. \\ &\quad \left. + \frac{\partial\Phi}{\partial\theta_x}\frac{\partial\theta_x}{\partial\theta_{z'}} + \frac{\partial\Phi}{\partial\theta_y}\frac{\partial\theta_y}{\partial\theta_{z'}} + \frac{\partial\Phi}{\partial\theta_z}\frac{\partial\theta_z}{\partial\theta_{z'}}\right) \\ &= -\omega_{z'}\frac{\mathrm{d}\Phi}{\mathrm{d}\theta_{z'}} = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}. \end{aligned} \tag{4}$$

Chao, Leon; Marangoni, Rafael; Seifert, Frank; El Haddad, Darine; Pratt, Jon R.; Newell, David B.; Schlamminger, Stephan. "The Design of an Instrument to Realize Small Torque at NIST." Paper presented at NCSLI (National Conference of Standards Laboratories International) Workshop & Symposium, Denver, CO, US. August 22, 2020 - August 27, 2020.

We will refer to measurement of the voltage induced by the rotor as the *angular mode* of operation.

Now suppose that the motion is halted and an electrical current I is passed through the same loop of wire generating a torque about the z' axis. We will refer to the measurement of this current as the *torque mode*.

Consider the energy E of the coil,

$$E = \Phi I \tag{5}$$

so that torque Γ about z' can be expressed as

$$\begin{aligned}\Gamma &= -\frac{\mathrm{d}E}{\mathrm{d}\theta_{z'}} \\ &= -I\left(\frac{\partial\Phi}{\partial x}\frac{\partial x}{\partial\theta_{z'}}+\frac{\partial\Phi}{\partial y}\frac{\partial y}{\partial\theta_{z'}}+\frac{\partial\Phi}{\partial z}\frac{\partial z}{\partial\theta_{z'}}\right. \\ &\quad\left.+\frac{\partial\Phi}{\partial\theta_x}\frac{\partial\theta_x}{\partial\theta_{z'}}+\frac{\partial\Phi}{\partial\theta_y}\frac{\partial\theta_y}{\partial\theta_{z'}}+\frac{\partial\Phi}{\partial\theta_z}\frac{\partial\theta_z}{\partial\theta_{z'}}\right) \\ &= -I\frac{\mathrm{d}\Phi}{\mathrm{d}\theta_{z'}}.\end{aligned} \tag{6}$$

If the system remains identical between the voltage measurement and the torque measurement, i.e. both measurements are taken at the same $\theta_{z'}$, then the partial derivatives of the the spacial coordinates, subsequently named the *calibration factor* denoted by ξ and equal to $-\mathrm{d}\Phi/\mathrm{d}\theta_{z'}$, will be eliminated when eqs. 4 and 6 are combined to become

$$\frac{V}{\omega_{z'}} = \frac{\Gamma}{I} \quad \text{and hence} \quad \Gamma = \frac{VI}{\omega_{z'}}. \tag{7}$$

Interestingly, this method of torque realization is independent of any lever arm measurement.

III. Design Overview

Torque wrenches and screwdrivers are often calibrated at the point of use with handheld commercial low torque reaction sensors. As an example, a series of five commercial devices are required to fully span from 0.01 N m to 1 N m and each have a relative uncertainty of 0.25% of its full measurement range [7]. The NIST Torque Instrument (NTI) should directly realize torque spanning the entire range of 0.01 N m to 1 N m with a relative uncertainty of 0.1% or better to offer a competitive alternative.

Two designs are analyzed in this manuscript. Design type X shown in fig. 1 utilizes a magnet system where the magnetization points along the X direction. Design type Z as shown in fig. 2 uses a magnet system where the magnetization points along the Z direction.

Each design consists of two identical permanent ring magnets oriented in attraction but vertically separated to form an air gap and concentrically fixed to a central shaft. The shaft rotates via a low friction bearing such as an air bearing. A D-shaped coil is fixed in space as shown in fig. 3, vertically centered in the air gap.

The magnet system is free to rotate relative to the coil for angular mode. Rotation can be achieved either by direct drive of a motor or the end user may provide a physical impulse. The rotation is sensed with a rotary encoder mounted centered on top of the upper magnet.

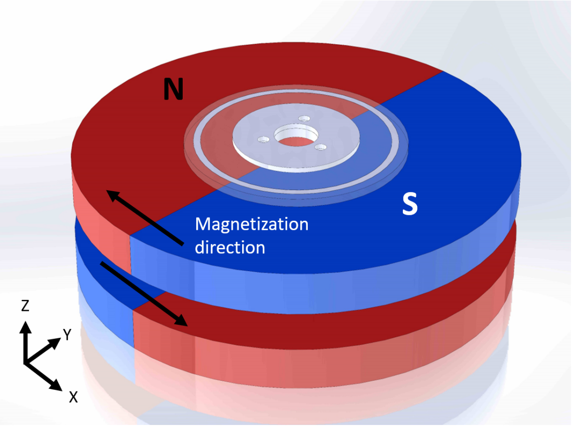


Fig. 1. Type X magnet configuration. Magnetization is along the X direction. Rotary optical encoder scale is mounted to the top of the upper magnet. Central shaft and bearing are hidden for clarity.

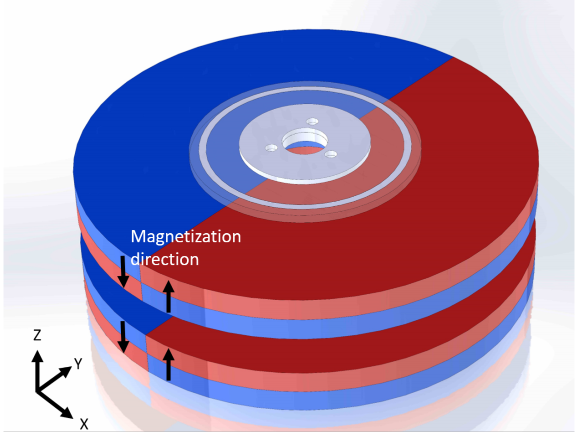


Fig. 2. Type Z magnet configuration. Magnetization is along the Z direction. Rotary optical encoder scale is mounted to the top of the upper magnet. Central shaft and bearing are hidden for clarity.

Like most Kibble-style instruments, measuring the calibration factor accurately in angular mode is the central challenge. Thus, we speculate an absolute determination of $\omega_{z'}$, or more specifically the change in rotational angle $\mathrm{d}\theta_{z'}$, will be the largest metrological hurdle. Insightful design of the magnet system can reduce the difficulty, as explored later in section V.

The vertical component of the magnetic flux density $B_{z'}$ in the air gap is plotted in figs. 4 and 5 for the Type X and Z magnet system, respectively. For the Type X magnet system, $B_{z'}$ is most dense near the outer edges furthest away from the split line. For the Type Z magnet, $B_{z'}$ is close to uniformly distributed throughout the air gap with a polarity difference between the two halves.

Based on the theory discussed in section II, the NTI functionality is neither affected by coil geometry nor ec-

Chao, Leon; Marangoni, Rafael; Seifert, Frank; El Haddad, Darine; Pratt, Jon R.; Newell, David B.; Schlamminger, Stephan. "The Design of an Instrument to Realize Small Torque at NIST." Paper presented at NCSLI (National Conference of Standards Laboratories International) Workshop & Symposium, Denver, CO, US. August 22, 2020 - August 27, 2020.

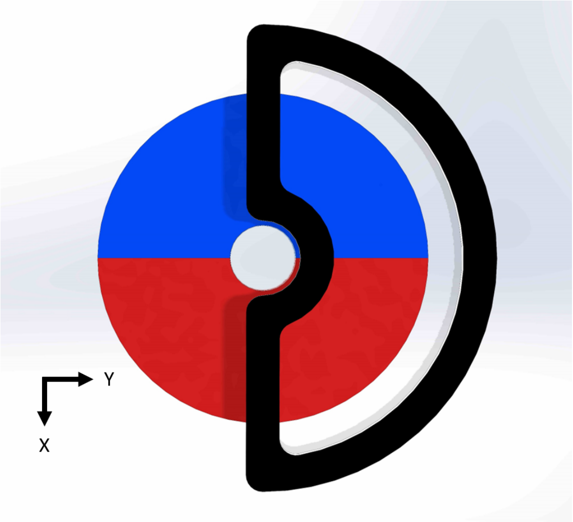


Fig. 3. Top view of magnet assembly where the upper magnet and encoder scale are hidden for clarity to show D-Coil. This orientation represents where $\theta_{z'} = 0$.

centric magnet motion, as long as these parameters do not vary between both angular and torque mode and the motion along the eccentric path is measurable. Furthermore, while the uniformity of the magnetic flux also does not play a role in the theoretical operation of the NTI, it does govern experimental aspects such as the sampling method for measuring the angular velocity.

In torque mode, we must minimize friction, magnetic, and gravitational forces in the central bearing to maximize sensitivity while maintaining single DOF motion. Additionally, it is still unclear how the torque wrench or torque screwdriver should interface with the NTI, especially if both are magnetic.

IV. Feasibility Calculations

To bound the design parameter space, we assume a maximum torque of $\Gamma = 1\,\mathrm{N\,m}$ and aim for an approximate size of 130 mm diameter × 36 mm height, similar to existing geometries of other calibration instruments in this range of torques.

A. Magnet System

For both Type X and Z, an NdFeB ring with geometry of 130 mm outer diameter × 12 mm height with a 25 mm inner bore will be used for this feasibility assessment. NdFeB was chosen for its high magnetic flux density and relatively high temperature coefficient of $10^{-3}\,\mathrm{K}^{-1}$. Half of the magnet has the opposite polarity as the other with a magnetic flux density magnitude of $\mid B \mid = 1\,\mathrm{T}$, a magnitude achievable by most magnet manufacturers. A second identical disk is concentrically mounted to the first, offset by 12 mm where the air gap is defined. Again, the magnetic flux density $B_{z'}$ on the midplane of the air gap was simulated for both Type X and Z in figs 4 - 5. The magnet is set to rotate at angular velocity $\omega_{z'} = 6\pi$ rad/s, three rotations per second, a reasonable speed for generating larger voltages between 1-10 V in the coil.

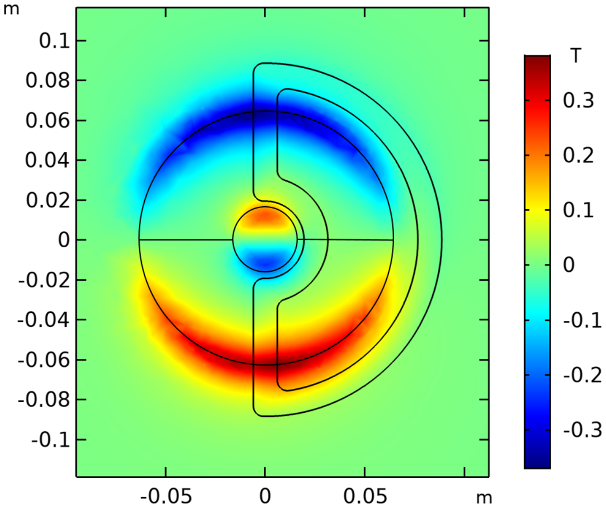


Fig. 4. Finite Element Analysis of the Type X magnetic flux density $B_{z'}$ distribution along the midplane of the air gap. Highest flux concentrations are along the edges of the magnet system furthest from the split line.

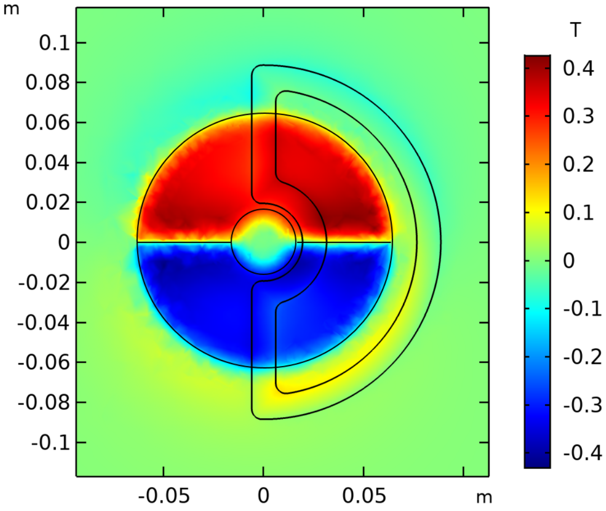


Fig. 5. Finite Element Analysis of the Type Z magnetic flux density $B_{z'}$ distribution along the midplane of the air gap. The flux concentration is more uniformly distributed throughout the air gap area.

B. Coil

The D-shaped coil will have a cross-sectional area of 12 mm × 6 mm, allowing for about 500 turns of AWG-26 wire, and an overall average radius r_e of 6.5 cm or an enclosed coil area of approximately $60\,\mathrm{cm}^2$, about half the footprint of one magnet. Covering half the magnet optimizes the sweep range for which maximum torque is generated. The maximum allowable current in AWG-26 wire is 2.2 A.

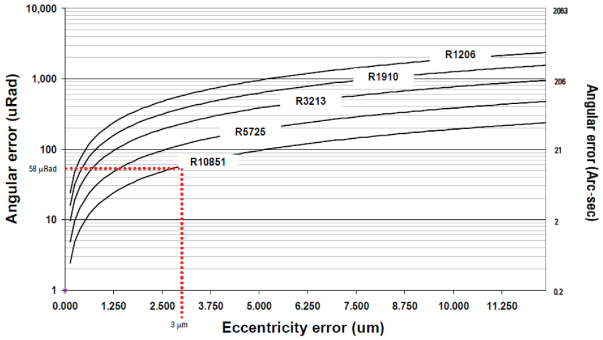


Fig. 6. The effect of eccentricity on angular accuracy over 2π radians as specified by the manufacturer, MicroE [8]. The rotary glass scale model R5725 will be used with a Mercury M1500 read head. For errors below $20\,\mu$rad, the eccentricity error on the concentricity alignmnent of the scale to the axis of rotation must be below $1\,\mu$m.

C. Rotary Encoder

In our baseline design, a microE R5725[1] rotary encoder scale will be rigidly mounted to the top surface of the top magnet and a microE Mercury M1500 read head was chosen for prototyping because this model is readily available for use by our group. The achievable uncertainty by this encoder is specified as 1 µrad when it is perfectly aligned, i.e., when the rotational run out of the scale is zero. Otherwise, the eccentricity of the trajectory causes an error over a full revolution indicated by fig. 6. Various alignment techniques can be implemented to reduce the eccentricity error but all require additional electronics.

D. Voltage, Current and Timing Measurement Hardware

It is assumed that the hardware capabilities of measuring voltage, current, and time to 0.1 % pale in comparison with that of the rotary encoder. General ideas regarding this topic are expounded upon in section V.

V. Measurement scheme

It is generally proper metrological practice to conduct measurements using an A-B-A scheme meaning a torque mode measurement is conducted both before and after an angular mode measurement. The methodology for measuring each mode is discussed.

A. Angular Mode

Recall from section II that the calibration factor ξ measured in angular mode is defined as

$$\xi = -\frac{\mathrm{d}\Phi}{\mathrm{d}\theta_{z'}} = \frac{V}{\omega_{z'}}, \tag{8}$$

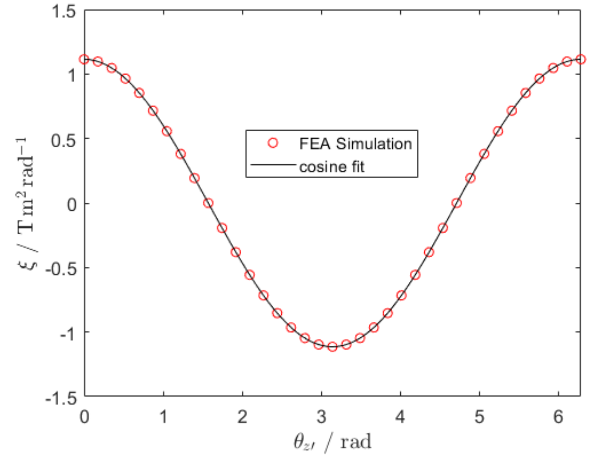


Fig. 7. Calibration factor ξ plotted over a full revolution of the Type X magnet. Torque extrema $(\mathrm{d}\xi/\mathrm{d}\theta_{z'} = 0)$ occur at $\theta_{z'} = n\pi$ where $n = 0, 1, 2, \ldots$. The points follow a cosine with an amplitude of $1.1\,\mathrm{T\,m^2\,rad^{-1}}$ (black line). A cosine fit can be applied to resolve the profile shape.

where the two variables, V and $\omega_{z'}$, must be determined to a relative uncertainty of 0.1% or better. Angular velocity $\omega_{z'}$ is measured by taking the change in rotational angle $\Delta\theta_{z'}$ over a measured time duration Δt.

However, the two different types of magnet systems proposed allow for two different methods of sampling $\Delta\theta_{z'}$. For the type X, the calibration factor is plotted over one revolution of the magnet in fig. 7 and follows a cosine. Hence, a high resolution angular measurement is not required. Instead, it can be substituted with a simple optical flag which is tripped once every revolution. $\Delta\theta_{z'}$ would have a baseline resolution of 2π rad and the Δt would be the time for one revolution or 0.33 s. The calibration factor could be interpolated with a cosine fit. The drawback to this method is that the true calibration factor profile must not have distortion greater than 0.1% from a cosine fit. If two identical magnets are unattainable, the cosine profile can be "tuned" to the proper shape with precision iron yokes or by allowing X-Y adjustability between the two magnets.

For the Type Z setup, the calibration factor is plotted over one revolution of the magnet with a calibration factor profile as shown in fig. 8. Here it is difficult to predict the profile shape a priori thus a cosine interpolation is invalid and rapid sampling of angular position and time must be achieved. Torque extrema where a wide region of $\mathrm{d}\xi/\mathrm{d}\theta_{z'} = 0$ occur at $\theta_{z'} = n\pi$ rad where $n = 0, 1, 2, \ldots$. At these angles, the change in ξ only exceeds 0.1% when $\theta_{z'} > n\pi \pm 50\,\mathrm{mrad}$.

Let's assume here that $\Delta\theta_{z'}$ needs to be sampled every 20 mrad to resolve the profile. Thus, the absolute angular error allowed for this measurement would be $20\,\mu$rad. Referencing fig. 8, the rotary scale eccentricity error must be less than 1 μm.

Given the magnet has an angular velocity of $\omega = 6\pi\,\mathrm{rad\,s^{-1}}$, the induced voltage would have a peak value on the order of 10 V making the voltage metrology of 0.1% relatively trivial

[1] Certain commercial equipment, instruments, and materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

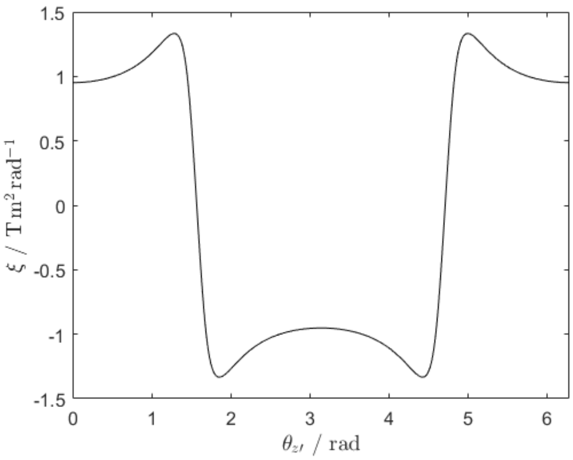


Fig. 8. Calibration factor ξ plotted over a full revolution of the Type Z magnet. Torque extrema where a wide region of $\mathrm{d}\xi/\mathrm{d}\theta_{z'} = 0$ occur at $\theta_{z'} = n\pi$ rad where $n = 0, 1, 2, ...$ High resolution sampling of $\omega_{z'}$ and V are necessary to resolve the profile shape.

compared to the angular metrology. For example, a Fluke 177 handheld multimeter already meets performance specifications with an accuracy of 0.09% for DC voltage measurements. Though such a device is lacking in terms of external triggering features and timing jitter on when the voltage is actually measured, it demonstrates that obtaining a sufficiently accurate DC voltage measurement is not difficult.

The same is true for the time metrology. For example, a low cost ublox LEA-M8F time and frequency reference can run independent of GPS with a 1×10^{-7} hold-over after 24 hours. Since a computer will be necessary for operating the NTI, the embedded computer clock provides another option that meets the timing measurement needs.

B. *Torque Mode*

Once ξ is determined in angular mode, the magnet is to be servoed to $\theta_{z'} = 0$ or π rad where the torque is at an extremum, i.e. $\mathrm{d}\xi/\mathrm{d}\theta_{z'} = 0$. With a controllable current source, a current I will be applied. The current will be measured either directly or via the equivalent voltage drop across a calibrated resistor. A torque wrench or screwdriver to be tested is then inserted into the NTI and I is increased until the tool clicks. It is critical to design the system such that the magnetic interaction between the tool and the NTI is minimized otherwise an additional source of error appears. This torque is measured as

$$\Gamma_{\mathrm{on}_1} = \xi I_{\mathrm{on}_1} \tag{9}$$

Next, a null measurement is conducted with the tool removed. Any spurious torques from bearing friction or gravity loading will be measured here as

$$\Gamma_{\mathrm{off}} = \xi I_{\mathrm{off}} \tag{10}$$

Finally, following the A-B-A measurement principle, a bracketing torque measurement is taken

$$\Gamma_{\mathrm{on}_2} = \xi I_{\mathrm{on}_2} \tag{11}$$

The torque generated by the NTI to calibrate the tool under testing is defined as

$$\Gamma_{\Delta} = \xi(\frac{I_{\mathrm{on}_1} + I_{\mathrm{on}_2}}{2} - I_{\mathrm{off}}) \tag{12}$$

By injecting the maximum allowable current 2.2 A into the coil at $\theta_{z'} = 0$, the Type Z magnet system can generate a peak torque of 2.1 N m, almost twice that of the Type X system of 1.1 N m due to its magnetization naturally pointing along the Z direction.

VI. Coil Optimization

It is interesting to take notice of the residual magnetic field residing outside of the air gap bounds. Examining figs. 4 - 5, a faint halo encircles the outer circumference of the magnet and has a $B_{z'}$ value of approximately 0.1 T or -0.1 T. In the Type X model, the sign of this halo $B_{z'}$ value matches that of its corresponding magnet edge and in the Type Z model, the sign of $B_{z'}$ is opposite of its corresponding magnet edge. This sheds insight on how to determine the optimal radius r_e of the coil to maximize the output torque.

Fig. 9 shows the maximum torque capability each system as a function of coil radius r_e at $\theta_{z'} = 0$. In the Type X system, there is no clear maximum since the flux density in the halo has no sign change, i.e. the optimal r_e is ∞ . Subjectively, one can choose r_e based on diminishing returns which most likely lies in the range of 8 - 9 cm. In the Type Z system, a clear maximum exists as the flux density in the halo goes through a sign change. Thus the coil radius where maximum torque can be attained is 6.7 cm.

The amount of power dissipation by the D coil at maximum current is on the order of 100 W, which makes the calibration factor to first order independent of thermal expansion but may present a safety concern to the user due to power density. This issue can be treated with thermal management techniques such as heat sinks or by modifying the parameters of the electromagnet, i.e. increasing $B_{z'}$, decreasing coil resistance, or by adding a second D-coil on the unused half of the electromagnet.

VII. Discussion

The two designs proposed have both advantages and practical limitations associated with fabrication and alignment of components. The advantage of Type X is that it is a sampling-free method since a perfect cosine ξ profile is assumed and verified by the induced voltage signature. A single optical flag can be implemented instead of a rotary encoder which is difficult to align. The caveat is the cosine profile must not have a distortion of more than 0.1% from a true cosine function. This is governed by the physical alignment and variation of magnetization between the two magnets and may prove difficult to achieve.

A possible remedy could be to add X and Y adjustability of one magnet relative to the other so that the cosine shape

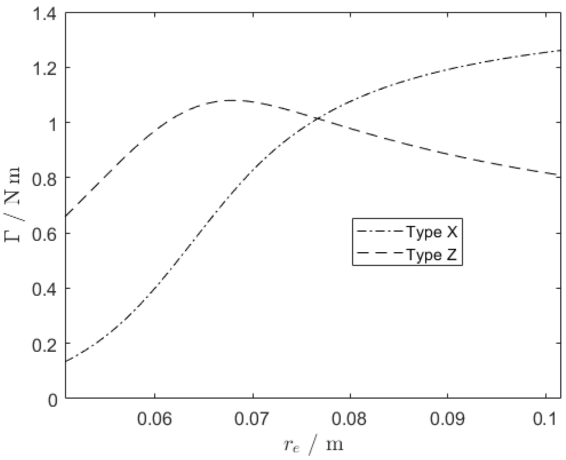


Fig. 9. Output torque with varying coil radii r_e for the Type X and Type Z magnets when at $\theta_{z'} = 0$. Maximum torque for Type Z is when $r_e = 6.7\,\text{cm}$.

can be empirically "tuned" by observing the induced voltage signal over multiple rotations.

The advantage and disadvantage of Type Z is opposite of to that of Type X. The level of alignment and variation of magnetization between the two magnets are relaxed since the entire ξ profile is to be scanned. The meticulous alignments required for Type X are shifted onto aligning the rotary encoder scale instead as the geometric center of the scale must be coincident with the axis of motion of the magnet system.

A possible remedy could be to widen the flat spot and increase $\Delta\theta_{z'}$ where $\xi < 0.1\%$, thereby lowering the error due to eccentricity of the encoder scale. This can be achieved by installing precision iron yokes to steer the magnetic flux to obtain a more uniformly distributed magnetic field in the air gap.

No matter which design, both require high levels of inter-component alignments and careful attention to ensure the magnetic interaction between the tool to be calibrated and the NTI is negligible. It seems wise to take a hybrid approach, perhaps using the encoder for a low resolution scan of the Type X magnet profile, relaxing the criteria for both a near-perfect encoder scale alignment and a near-perfect cosine interpolation.

VIII. Conclusion

A theoretical feasibility study was conducted at NIST for a method to directly realize small torques for calibrating torque wrenches and torque screwdrivers geared towards end users and the results show promise. A maximum torque of approximately $1.1\,\text{N m}$ can be achieved with both the Type X and Type Z designs. The next steps include the detail designing of both types of magnet systems and procure the components necessary for prototyping. The prototypes will be measured against one another to fundamentally determine which model is superior. Further research will be required for ultimately evolving the prototype into a commercial product and will depend on collaborating with instrument manufacturers.

References

[1] ISO 6789-2, Assembly tools for screws and nuts — Hand torque tools — Part 2: Requirements for calibration and determination of measurement uncertainty, 2017.

[2] The History of Morehouse, https://www.mhforce.com/Home/AboutUs?tab=History

[3] A Knott "Developments at NPL to improve calibration of force and torque," *Measurement + Control*, vol. 38/7, 246-251, 2004.

[4] A Nishino *et al* "Design of a new torque standard machine based on a torque generation method using electromagnetic force," *Meas. Sci. Technol.* 28 025005, 2017.

[5] D Haddad *et al* "Measurement of the Planck constant at the National Institute of Standards and Technology from 2015 to 2017," *Metrologia*, vol. 54, 633-641, 2017.

[6] B P Kibble *et al* "Principles of a new generation of simplified and accurate watt balances," *Metrologia*, vol.51, 132-139, 2014.

[7] Mountz MTX series low torque reaction sensors, https://www.mountztorque.com/Product-Type/Torque-Analyzers-and-Sensors/Torque-Sensors/Torque-Sensors-MTX

[8] Celera Motion, technical paper on rotary scale alignment, https://www.celeramotion.com/microe/support/technical-papers/rotary-scales-alignment/

The UCEF Approach to Tool Integration for HLA Co-Simulations

Thomas Roth, Christopher Lemieux and Martin Burns
National Institute of Standards and Technology
Gaithersburg, Maryland, U.S.
thomas.roth@nist.gov, christopher.lemieux@nist.gov, martin.burns@nist.gov

***Abstract*—Cyber-Physical Systems (CPS) are complex systems that require expertise from multiple domains in their design, implementation, and validation. One cost-effective technique for validation of CPS is the integration of two or more domain-specific simulators into a joint simulation called a co-simulation. Standards such as the High Level Architecture (HLA) have been developed in part to simplify the co-simulation development process. However, CPS co-simulation still requires significant expertise, especially when the goal is the integration of a new domain-specific tool or simulator. The U.S. National Institute of Standards and Technology (NIST) has released a software platform called the Universal CPS Environment for Federation (UCEF) to simplify the development of CPS co-simulations. UCEF provides two approaches to integrate tools and simulators. The first approach is a Java library called the UCEF Gateway that limits the development effort to a list of callback functions in a well-defined simulation life cycle. The second approach is a Representational State Transfer (REST) server developed using the gateway for applications that can implement a Transmission Control Protocol (TCP)/Internet Protocol (IP) client. This paper describes how both approaches are implemented to expedite the integration of new domain-specific tools and simulators.**

***Keywords*-application programming interface, co-simulation, cyber-physical systems, high level architecture, tool integration**

I. Introduction

The design and implementation of Cyber-Physical Systems (CPS) requires significant expertise from multiple domains to ensure smooth operation. For a more formal definition, CPS consist of devices that use logical computation informed by measurements of the environment to actuate physical changes. CPS are common in critical infrastructure such as smart manufacturing, autonomous vehicles, and smart grid [1]. Failure of these systems has great economic and social costs, and validation is required to minimize the risk of failure prior to deployment. However, deployed CPS can be larger than city-scale and it is impractical to prototype all design decisions due to the immense cost associated with deployment. One cost-effective validation technique to overcome this challenge is co-simulation. Co-simulation is the integration of multiple domain-specific simulators into a common execution environment to produce results that more closely resemble the deployed system. The integration of existing simulators is a more scalable solution than the development of new, more complex simulators that can model all CPS dynamics in a single environment. It is quite common in smart grid research, for instance, to perform co-simulation that integrates a network simulator with a power system simulator [2].

The IEEE 1516-2010 High Level Architecture (HLA) is one standard for the co-simulation of distributed processes [3]. A single simulator or process is defined as a federate, and the collection of interacting federates is defined as a federation. The federation communicates and coordinates over middleware called a Runtime Infrastructure (RTI) which can be thought of as a shared message bus. The RTI provides a set of standardized services to the participating federates to facilitate the co-simulation. HLA was designed to be comprehensive and defines all the services that could be useful in distributed simulation whether or not those services are frequently used.

While HLA provides a rich and complete service set [4], the standard is complex and has features that may see minimal use in practical applications. Of the more frequently used HLA services, several can be implemented the same across all federates regardless of the domain or objective of any given experiment. But there is little publicly available information on which services are frequently used, and little guidance on how the services could be implemented to be reusable in a wide range of use cases. None of the services defined in the standard are labeled as optional, and it is not clear which parts of the standard must be implemented and which parts can be safely ignored. From the authors' experiences, learning HLA is a significant many-month process that does not greatly simplify the challenges of co-simulation.

The U.S. National Institute of Standards and Technology (NIST) is one of many groups that are developing software tools to reduce the burden of co-simulation development. The NIST tool, the Universal CPS Environment for Federation (UCEF), was released as a virtual machine that provides code generation of the HLA services for different simulators based on simple user-designed models [5]. One goal of UCEF is to provide a portable development environment where users can develop co-simulations without a background in distributed computing and the HLA standard. However, UCEF is only as powerful as the number of its supported simulators and the ease at which new simulators can be integrated.

This paper presents the approach to tool integration in UCEF. This approach makes assumptions on the HLA service set — in particular, it assumes that most of the services are not used — and uses those assumptions to produce a checklist of functions that must be implemented to integrate new simu-

Roth, Thomas; Lemieux, Christopher; Burns, Martin. "The UCEF Approach to Tool Integration for HLA Co-Simulations." Paper presented at Workshop on Design Automation for CPS and IoT, Sydney, AU. April 21, 2020 - April 21, 2020.

lators. Two methods for tool integration are presented: a Java library that can be extended to implement a new federate type, and a Representational State Transfer (REST) Application Programming Interface (API). These methods have been used to integrate several smart grid simulators into UCEF, and were developed out of a need to support software developers with no prior co-simulation experience.

The remainder of the paper is organized as follows. Section II provides an overview of related research into simplification of the co-simulation development process. Section III presents a brief overview of the UCEF software platform. The first approach to tool integration using a Java library is presented in Section IV, and the second approach using a REST API is presented in Section V. The paper is then concluded with Section VI.

II. Related Work

Several software platforms have been created to accelerate the HLA development process. These platforms let users model a CPS using a Domain Specific Modeling Language (DSML) and leverage code generation to transform user models into code that executes a subset of the HLA services. For instance, a user might define the input and output requirements of a federate in a table, and the software platform could then transform that table into skeletal code that requires minimal implementation from the user. At the forefront of these software platforms are the commercial design tools released by different HLA RTI vendors to simplify the use of their products [6][7]. While these tools are compatible with other RTI implementations, the HLA research community has attempted to develop open-source alternatives that perform similar functions due to concerns over cost.

One of the earlier open-source software platforms for HLA was the Command and Control Wind Tunnel (C2WT) produced from the Institute for Software Integrated Systems at Vanderbilt University. C2WT uses extensions to a graphical modeling environment called the Generic Modeling Environment (GME) to support the modeling and code generation of HLA federations [8]. Because GME was a desktop application that could only be accessed by one user, it was difficult to use in organizations that required model sharing between collaborators, and therefore a web-based variant was developed called the Web-based Generic Modeling Environment (WebGME) [9]. Vanderbilt University updated C2WT to use WebGME in a new software platform called the Cyber-Physical Systems Wind Tunnel (CPSWT). Public instances of CPSWT are hosted in the cloud at Vanderbilt University and the source code is available online through their GitHub repository [10]. NIST collaborated with Vanderbilt University to produce an offline version of CPSWT with additional support for several smart grid simulators. NIST released this software platform as an Ubuntu virtual machine called UCEF [5]. Other similar approaches have used Systems Modeling Language (SysML) diagrams to generate C++ executables compatible with Simulink models [11], and extensions to Eclipse that incorporate a DSML for HLA that can generate code for C++ federates [12]. All these software platforms attempt to minimize the implementation burden on the user by making assumptions on the default implementations of certain HLA services and using code generation.

Another significant contribution to the HLA open-source community is the HLA Development Kit Framework (DKF) [13]. The DKF is not a software platform, but an open-source Java library based on Java annotations. It provides a basic class structure that can be extended through inheritance to implement Java federates, and provides default implementations of most HLA services in parent and helper classes. In addition to the Java source code, the DKF is packaged with examples, tutorials, and documentation for the creation of federates using its simplified federate life cycle.

Another example that defines a federate life cycle with default implementations of the HLA can be found in [14]. This is not a software platform, nor a reusable library, but an example implementation of one federation using a well-defined life cycle. It defines a modular Federation Object Model (FOM) for data exchange between independent systems and prescribes a specific life cycle for federates in the form of state machines.

The Functional Mockup Interface (FMI) is a more recent co-simulation standard run as a Modelica Association Project [15]. While HLA attempts to define the complete set of services that could be useful in a distributed simulation, including services not commonly used in practice, FMI takes the opposite approach of trying to define the minimal set of functions required for co-simulation. FMI research efforts face similar challenges in trying to make the standard more accessible to users without deep knowledge of co-simulation. There are tool chains that use SysML models and code generation to automate portions of the federation development process [16], and there is a C++ library that can be leveraged to provide default implementations for most of the FMI functions [17].

In all of these cases, the goal has been to abstract the full range of HLA services and FMI functions that are visible to the end user, and provide default implementations for the set of services and functions that are federate independent. The reduced service set can then be considered as an API which minimizes user interaction with the standards documents. As a consequence, all the implementations are incompatible as they redefine in different ways the standardized services and function sets to improve user accessibility. The remainder of this document describes how the HLA services were redefined for the UCEF software platform.

III. Universal CPS Environment for Federation

The software efforts to simplify federation development share several common elements: they are often open-source projects that use some graphical language that leverages code generation to transform user models into federate code. However, there is often an assumption that it’s easy - or even possible - for a user to install and configure the software environment. Information technology (IT) policies such as firewall rules can prevent activities such as the installation of

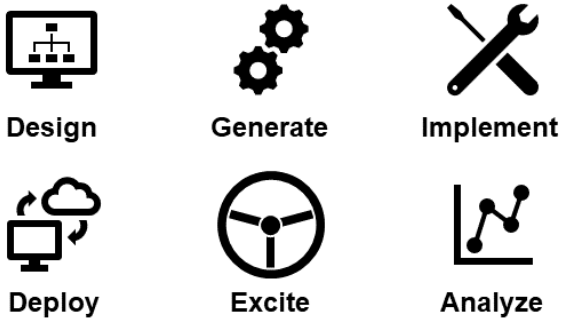


Fig. 1. The stages of Federate Development (top row) and Federation Deployment (bottom row) in the UCEF workflow

new software or access to a user account with administrative rights. These policies can make it very challenging to use some of the solutions mentioned in the related work.

This section provides an overview of the NIST software platform UCEF [5]. UCEF is a portable development environment created to expedite the development of HLA federates and federations. The main feature that distinguishes UCEF from similar approaches is its distribution as a self-contained virtual machine. This makes UCEF non-intrusive, easy to redistribute, and easy to install. It is distributed as an Ubuntu 16.04 virtual machine that runs a local WebGME server. The WebGME front end provides a graphical web environment where users can model federations using simple building blocks, and the back end uses JavaScript plugins that transform these models into stub code for different simulators. The current version of UCEF supports several simulators in the smart grid domain that include GridLAB-D, TRNSYS, and LabVIEW with additional support for native Java and C++ applications. The ease of development in UCEF rises from the separation of a federate implementation into two layers: a user layer which implements the federate behavior, and an infrastructure layer generated from WebGME that provides the default implementations for most HLA services.

Figure 1 shows the stages from federate development to federation deployment in the UCEF workflow. The current version of UCEF implements the top row related to federate development which consists of the stages Design, Generate, and Implement. These three stages produce an executable piece of software that can be run on any compute environment ranging from the UCEF virtual machine, to a desktop computer, to a node in the cloud. While UCEF also generates some simple bash scripts for deployment, the bottom row on federation deployment is still under development and the three stages of Deploy, Excite, and Analyze are notional.

The *Design* stage uses WebGME with the HLA metalanguage produced at Vanderbilt University for their platform CPSWT. A user produces a graphical model of a federate in a web browser which includes the specification of its simulator type (such as LabVIEW or Java program) and its various inputs and outputs. Figure 2 shows an example of a simple federate designed in this environment that both subscribes to and publishes one HLA message. Because WebGME is a web-application running on the local virtual machine, this modeling phase does not require an Internet connection despite the use of an web browser.

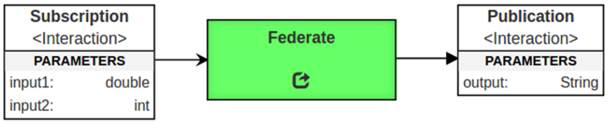


Fig. 2. An example WebGME Federate Model

The *Generate* stage is initiated when the user clicks a run button in WebGME to execute its code generation plugins on the federate model. WebGME plugins are written in JavaScript and use Embedded JavaScript Templates (EJS) to define the artifacts that should be generated for each of the supported simulator types. All artifacts produced from WebGME in UCEF are output as Apache Maven projects, regardless of whether they contain Java code. These artifacts have dependencies on the open-source Java RTI Portico and will not work with other RTI implementations [18].

The *Implement* stage varies dependent on the type of federate that was designed and generated, and may occur outside of the UCEF virtual machine. Appropriate domain-specific tools are used to implement each federate type, so Java files are implemented in Eclipse and LabVIEW projects are implemented in LabVIEW. In the cases that require an active license, such as LabVIEW, the generated files will have to be moved to a licensed machine to complete the federate development process.

Figure 3 shows how UCEF implements these three stages of the federate development process. The UCEF virtual machine contains a local WebGME server that is preconfigured with support for various types of simulators. When a user finishes the design and generate stage, stub code for the modeled federates is available outside of the UCEF virtual machine that can connect to an RTI and participate in a federation execution without any additional user implementation. However, the stub federate code contains no behavior and must be implemented by the user to fulfill its design goal. All the simulators shown in Figure 3, in addition to native Java and C++ applications, have been integrated into UCEF using the two approaches described in this paper.

The bottom row of Figure 1 on federation deployment is a notional representation of how deployment could work in UCEF. A federate designed in UCEF can be removed from the virtual machine and deployed in any environment, from a laptop to the cloud. Therefore, the mechanical process of deployment depends on an infrastructure that was provisioned and configured independent of the UCEF virtual machine, and this process may have significant differences from one work environment to another. There are, however, general deployment activities where UCEF could provide useful tools to expedite the deployment process.

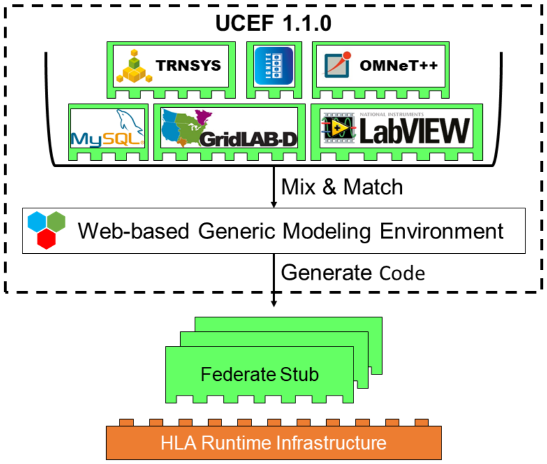


Fig. 3. Federate Development in UCEF

The federation *Deploy* stage needs to package the federates and their dependencies for deployment. First, the user needs to select the federates that should be deployed. It's possible that multiple instances of the same source code could be deployed as different federates in the same federation, and the selection process should handle this case. Then, the artifacts that contain the federates and their dependencies need to be collected from some database and packaged for deployment.

Once deployed, federates will be either configured or driven to execute a desirable scenario. The *Excite* stage perturbs the federation execution using a combination of static configuration files and dynamic runtime messages. For this stage, a simple scripting language could be incorporated into WebGME to generate configuration files that script the runtime behavior of federates. In addition, a suite of federates that enable user interaction – using either a graphical interface or a web server – could be developed and packaged with UCEF.

The *Analyze* phase involves both runtime monitoring of the federation and the use of database storage for offline analysis. For both cases, analytic federates are required to either allow user interaction at runtime or interface with different database systems for data logging.

The current version of UCEF supports the generation of bash scripts to run the federates in the virtual machine, and includes a database federate to store the results of the federation execution. The remainder of this paper focuses on the first row related to federate development and discusses how code generation, and default HLA service implementations, have been used to simplify the federate development process in UCEF.

IV. UCEF Gateway

The first approach to ease federate development is an open-source Java library called the UCEF Gateway that implements a simplified federate life cycle [19]. The gateway implements a main loop that yields control to user-implemented callback functions at specific points in this life cycle. The gateway was developed based on the following three requirements:

1) usable without HLA expertise
2) easy to integrate new things
3) agnostic to the federation data model

To satisfy the first two requirements, the gateway does not support the following HLA services: federation save, federation restore, ownership management services, and data distribution management services. A UCEF federation executes one experiment from start to finish and then terminates, without federates joining or leaving during the federation execution. Therefore, there is no need to load a prior state or handle the distribution of object instances due to sudden changes in the federation membership. The data distribution management services are useful for improved scalability, but the HLA implementation of regions as an unsigned integer is unwieldy. Each federate implementation must have the same region encoder and decoder functions to have consistent interpretations of the integer value, and this creates a new scalability problem due to the difficulty of configuration management. It's better to address scalability using traditional networking approaches rather than the use of HLA regions [20].

The third requirement distinguishes the UCEF Gateway from the HLA DKF which requires explicit Java annotations for declared variables that represent federation data. This requirement was derived from the need for a mechanism to ease the integration of entire simulators, not individual simulations. A simulator requires one reusable federate implementation that can support any simulation with an arbitrary data model. An approach that requires specification of a data model will need additional user implementation whenever a federate is integrated into a new domain or scenario, which defeats the purpose of a gateway library.

Since its release, the UCEF Gateway has been used to integrate several simulators: GridLAB-D, TRNSYS, and LabVIEW. Based on the lessons learned from these applications, a revised version with a modified federate life cycle has been released. This section summarizes the UCEF Gateway and highlights the modifications since its original publication.

A. Time Management Strategy

The gateway executes a well-defined life cycle with callbacks to the user application that can be used to define federate behavior. During the callbacks, the user code can use the public methods of the gateway library to perform functions such as sending data to the federation and querying the FOM. The gateway defines the time management strategy on behalf of the user application, and this strategy cannot be modified. All gateway implementations are both time constrained and time regulating to operate in lockstep with federation logical time. Logical time progression uses the HLA time advance request service with a fixed step size for the duration of the federation execution. The logical step size can be configured by the user in the gateway configuration files. At this time, the next event request service is not supported.

Roth, Thomas; Lemieux, Christopher; Burns, Martin. "The UCEF Approach to Tool Integration for HLA Co-Simulations." Paper presented at Workshop on Design Automation for CPS and IoT, Sydney, AU. April 21, 2020 - April 21, 2020.

During its life cycle, the gateway assumes the federation has three synchronization points: ready to populate, ready to run, and ready to resign. The gateway assumes that another federate registers these synchronization points, and that federate also determines when the federation synchronizes on each point. In UCEF this federate is called the federation manager, and it gates progression through the federate life cycle by delaying its synchronization until it determines the federation is ready to progress. These synchronization points divide the federate life cycle into three distinct stages: initialization, logical time progression, and termination.

B. Federate Life Cycle

Figure 4 shows the UCEF Gateway federate life cycle. The rectangles are the user-implemented callback functions. Several transitions between callbacks depend on federation synchronization events, which are indicated with a labeled dotted line below the transition. A gateway implementation can block on the first transition, labeled *JoinFederation*, when it tries to join a federation that has not yet been created. The other synchronization events correspond to the three synchronization points discussed in the previous subsection on time management. In addition, the transition labeled *time advance grant* blocks until the federation as a whole advances its logical time to the next logical time step.

Table I describes each callback function in the life cycle. Four of the callback functions that begin with the word *receive* are consolidated in Figure 4 as the single state *receive data*. The order of these callbacks is arbitrary and interleaved, as it depends on the RTI implementation. It is possible that some of the *receive* callbacks do not occur for a given logical time step, and that the callbacks occur in different orders between logical time steps. However, all the *receive* callbacks will be handled prior to the gateway invoking the *step* callback.

The life cycle and callback functions do not show the public methods available in the gateway library for interaction with the federation. Some of these methods are intuitive, such as sending data to the federation and querying for the data type of received data. These methods are summarized in the original gateway publication, and unchanged in the revised version. Of note is that the gateway library provides a polling mechanism to receive new data that can be executed anywhere in the life cycle except *before join federation* and *before exit*. Although Figure 4 seems to indicate that data arrives in one bulk read operation each logical time step, the user application could choose to poll for data where it is needed. The two noted exceptions are merely because the gateway does not exist as a federate in a federation for those two stages of the life cycle, so the poll data operation is undefined.

A few of the callback functions are new since the original release of the UCEF Gateway. The callback *before join federation* was added to give a user a proper initialization method in the life cycle rather than relying on the Java constructor to serve this role. The *before first step* and *before ready to resign* callbacks were added to give unique meaning to the first and last logical time step of a simulation. Before these callbacks were introduced, conditional logic had to be inserted into the *step* callback to determine whether a time step was an intermediate or edge step. This complicated the user code more than the addition of two optional callbacks that could be ignored when the first and last steps are not distinguished. Likewise, the *receive object registration* and *receive object deleted* callback were added to prevent the use of conditional logic inside the *receive attribute reflection* callback when processing an object instance for the first time. While this has led to the introduction of five additional callback functions, the default behavior for each callback is a no-operation.

V. REST API

The UCEF Gateway was intended to simplify federate development by providing reasonable default implementations for the HLA services that did not change between federate implementations. However, in practice, most gateway applications that integrated simulators into UCEF used a simple client-server architecture with TCP/IP sockets. It is much easier to embed a socket in a simulator to pump its data to a server than extend the simulator source code to implement the gateway callback functions. This common use of the gateway led to redundant socket code between gateway implementations — a problem the gateway was designed to alleviate — and the creation of unique communication protocols for each simulator. This section describes the first attempt at a REST API built using the UCEF Gateway to provide a common server implementation for these TCP/IP socket applications. This REST API is available as a standalone federate distributed with UCEF that can be incorporated into any federation.

The reusable TCP/IP server was implemented as a REST API rather than a custom socket protocol for three reasons. First, the client code would be shielded from the potentially long synchronization delays caused by HLA logical time progression. A sensor or small Internet of Things (IoT) device might want to produce a stream of data at a constant frequency and avoid blocking calls. The fast response of a REST implementation will support these devices without the need for a multi-threaded implementation. Second, the REST implementation eliminates the need for constant heartbeat messages between the client and server to ensure a persistent socket connection. This reduces the overhead for the client implementation. Third, the REST API introduces another layer of abstraction that may make it easier for users to develop new federates without knowledge of the HLA standard.

The HLA standard already defines a REST API for interacting with the RTI [21][22]. It is important to note that the standard API exposes the complete set of RTI services, which includes the services that are rarely utilized and the services where default implementations are sufficient for most use cases. Use of the standard API represents a significant implementation burden on the user, and loses the benefits of a simplified approach like the UCEF Gateway. For this reason, a new REST API implemented on top of the UCEF Gateway was developed and is presented in this section.

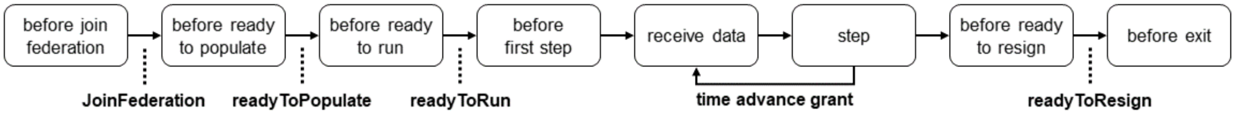


Fig. 4. UCEF Gateway Life Cycle

TABLE I
GATEWAY CALLBACK FUNCTIONS

Callback	Description
before join federation	Perform basic initialization that is HLA independent, such as initialization of data structures
before ready to populate	Perform initialization that requires a joined federation, such as registration of object instances
before ready to run	Perform initialization that requires other federates, such as the exchange of initial values
before first step	Perform one-time actions for the first logical time step, such as starting a simulation
receive object registration	Handle a discovered object instance
receive attribute reflection	Handle an attribute reflection for one discovered object instance
receive object deleted	Handle a removed object instance
receive interaction	Handle one received interaction
step	Perform the logic executed each logical time step, such as updating object attributes
before ready to resign	Perform one-time actions for the last logical time step, such as the exchange of final values
before exit	Perform any cleanup, such as stopping the simulation and closing output files

TABLE II
LIST OF ENDPOINTS

Endpoint	Method	Request Format	Response Format
/status	GET	(none)	FederateStatus
/join	POST	ClientPost	FederateStatus
/dostep	POST	ClientPost	FederateStatus
/ping	POST	(none)	200 OK

A. Endpoints

Table II lists the endpoints defined for the REST API. The request and response formats indicated in this table are defined in the following subsection on payloads. The *ping* endpoint is intended to be a light-weight heartbeat message to check the status of the server. When the federation is starting, or when the client loses connection with the server, the *ping* message can be used to periodically check if the server is online. The *status* endpoint returns the current state of the server. Because the server implements the UCEF Gateway, it represents a federate in some federation, and its status contains information such as the current logical time and the most recent values for data exchanged in the federation. The *join* endpoint is used to tell the server that the client wants to join the federation, and contains some details about the client's identity and data model. The *dostep* endpoint is used to tell the server that the client is ready to advance to the next logical time step, and contains the set of client data that should be broadcast to the federation.

B. Payloads

JavaScript Object Notation (JSON) is used to define all the payload formats. The endpoints define two payloads called *ClientPost* and *FederateStatus*. However, these payloads contain HLA interactions and object instances. This subsection will first define the JSON format for these primitive HLA data structures, and then the formats used in Table II.

Listing 1 defines the JSON for object instances. An object in HLA can be thought of like a variable in a programming language. The user defines a data type (the object class) and a unique variable name (the instance name). The object class determines the structure of the data associated with the object instance in much the same way the data type determines the structure of a variable. The *classPath* field is the fully qualified path for the object class, and the *instanceName* field is the unique identifier of a particular object instance. The *attributes* array is a list of name-value pairs for the data associated with the specified object instance. All of the attribute values are encoded as strings, although the attribute could be any of the data types defined in the FOM. In the current implementation, the client and server are both pre-configured with the FOM so both sides can convert the string value into the correct data type. However, future work will have the client send the FOM to the server by embedding it into the payload of the *join* endpoint.

Listing 1
OBJECTINSTANCE JSON FORMAT

```
{
  "classPath": "ObjectRoot.ClassName",
  "instanceName": "ObjectInstanceName",
  "attributes": [
    {
      "name": "attrName",
      "value": "asString"
    }
  ]
}
```

The interpretation of Listing 1 depends on the direction of data flow. When the client sends an object instance to the server, it is a request to publish updated values for an object instance registered by the client. When the server sends an object instance to the client, it is a notification that the values for that object instance have changed in the federation.

Listing 2 shows the JSON format for interactions which is almost identical to the format for object instances. Because interaction instances are not assigned unique identifiers, as they have no persistent state that changes over logical time, the *instanceName* field has been dropped. In addition, to conform with the HLA naming conventions, the *attributes* field has been renamed to *parameters*. Otherwise, the format and use of this JSON payload is identical to object instances.

Listing 2
INTERACTION JSON FORMAT

```
{
  "classPath": "InteractionRoot.Class",
  "parameters": [
    {
      "name": "paramName",
      "value": "asString"
    }
  ]
}
```

Listing 3 shows the *ClientPost* JSON format that is sent from the client to the server. The client provides its internal state, represented with three boolean variables, as well as the list of interactions and object instances that should be published to the federation. *isJoining* is a boolean flag that indicates the client is ready to start the simulation, and *isLeaving* is a boolean flag that indicates the client is ready to stop the simulation. *isAdvancing* is a boolean flag that indicates the client has finished its current logical time step and is ready to advance logical time to the next iteration of execution. Not all permutations of values for these flags are valid, as a client cannot simultaneously join and leave. The valid permutations will be elaborated on in the following subsection on the state machine. The *interactions* and *objects* arrays contain updated values for all the interactions and object instances for the current logical time step. If an object instance has not changed from the previous time step, it can be omitted entirely from this array.

Listing 3
CLIENTPOST JSON FORMAT

```
{
  "isJoining": true/false,
  "isAdvancing": true/false,
  "isLeaving": true/false,
  "interactions": [ Interaction ],
  "objects": [ ObjectInstance ]
}
```

Of note is that the *ClientPost* payload provides no mechanism for the client to inform the server that it wants to register a new object instance. The server maintains a list of names for the client's registered object instances. When the client sends an object instance with an unknown *instanceName*, the server will automate the object registration process and update its internal list.

Listing 4 shows the *FederateStatus* JSON format that is sent from the server to the client. The server also provides its internal state using three boolean variables that will be further explained in the subsection on the state machine. *isSimulationActive* indicates the federation exists, and is either preparing to begin or has already begun logical time progression. *isDoStep* indicates the server is idle waiting on input from the client before it proceeds to the next logical time step. *isTerminating* indicates the simulation is over, and the client needs to prepare to exit. An additional *timeStep* value is provided to notify the client of the current logical time in the HLA federation. The *interactions* and *updatedObjects* arrays contain updated values received from the other federates. If an object instance has not been updated since the last status report, it is omitted entirely from this array. The *FederateStatus* JSON also contains a *newObjects* array that lists all the object instances that were registered by other federates since the last status update. If an object instance appears in the *newObjects* array for a given status report, it will not appear in the *updatedObjects* array.

Listing 4
FEDERATESTATUS JSON FORMAT

```
{
  "isSimulationActive": true/false,
  "isDoStep": true/false,
  "isTerminating": true/false,
  "timeStep": 0.0,
  "interactions": [ Interaction ],
  "newObjects": [ ObjectInstance ],
  "updatedObjects": [ ObjectInstance ]
}
```

C. State Machine

Figure 5 shows the state machine implemented by the REST server. This state machine does not show the exceptions that might cause the server to respond to the client with an error code. An exception would occur whenever the server receives a *join* or *dostep* request from the client in a state where there is no explicit transition labeled with that endpoint. The state transitions mirror the synchronization points identified in the gateway life cycle from Figure 4. The same labels are used for *JoinFederation*, *readyToPopulate*, *readyToRun*, *time advance grant*, and *readyToResign*. Three additional labels appear in the state machine transitions: two of the REST endpoints for *join* and *dostep*, and one transition labeled *exit condition*. Because the *ping* and *status* endpoints are valid in all states, they are not shown in the state machine.

The server has an internal state represented by the three boolean flags *isSimulationActive*, *isDoStep*, and *isTerminating*. These booleans have well-defined values for each state in the state machine as indicated in Figure 5. When the server produces a *FederateStatus* payload in response to a client request, the values of these booleans associated with the current

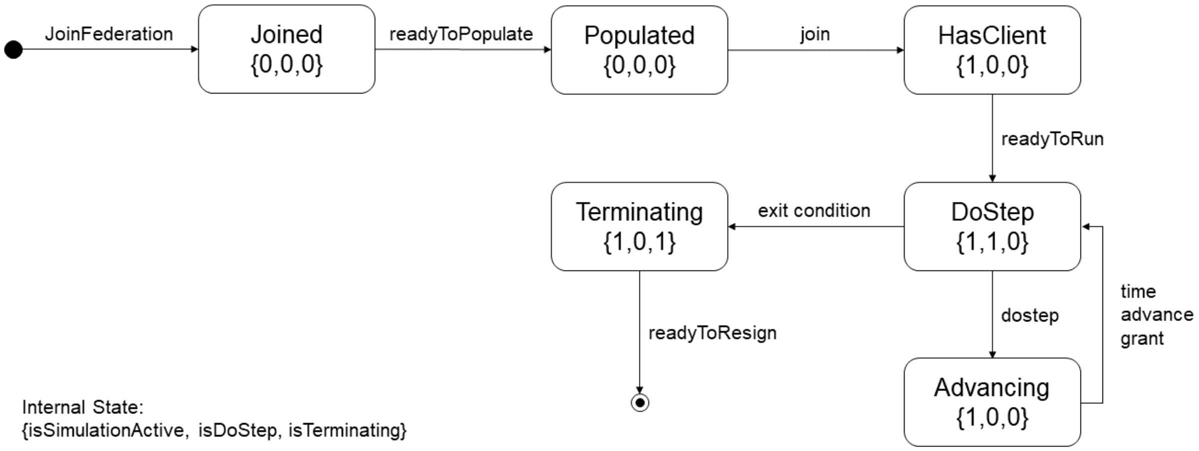


Fig. 5. State Machine of the REST Server

state are used to populate the contents of that payload. The remaining information is retrieved from the HLA federation.

VI. Conclusion

This paper provided a brief overview of UCEF and described two different approaches to tool integration for the platform. The first approach, an open-source Java library called the UCEF Gateway, is a mature implementation that has been used to integrate several new simulators into UCEF by different developers. The second approach, a REST server built using the gateway library, is a more recent development undergoing continuous refinement. Both approaches seek to simplify the tool integration challenge by providing default implementations for HLA services and exposing a reduced API that is more accessible for developers without co-simulation experience.

There are future plans to expand the REST server to support multiple simultaneous client sessions with the goal of creating something akin to an IoT gateway that integrates a large number of devices into a federation as a single federate. The server state machine will have to be updated to support multiple clients, and a session identifier will have to be inserted into the payloads so clients can be identified between calls to the different endpoints. During this process, it is likely the list of endpoints and the payload structure will undergo continuous refinement as the software matures.

One interesting research direction for this work is the performance characterization of different types of user applications using the two approaches. The REST server relies on socket communication rather than direct function calls and should be slower and more prone to bottlenecks than native UCEF Gateway implementations. However, the communication pattern of the user application — such as the frequency of communication and the size of the payloads — could result in vastly different performance profiles. It is likely that some types of user applications are ill-suited to using the REST server, while other types notice little to no performance degradation over a native gateway implementation. An investigation of these different performance characteristics would add another dimension beyond ease-of-use that must be considered when choosing to integrate a tool using one of the approaches.

Acknowledgment

Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

References

[1] E. R. Griffor, C. Greer, D. A. Wollman, and M. J. Burns, "Framework for cyber-physical systems: Volume 1, overview," Tech. Rep., 2017, doi: 10.6028/NIST.SP.1500-201.

[2] S. C. Müller, H. Georg, J. J. Nutaro, E. Widl, Y. Deng, P. Palensky, M. U. Awais, M. Chenine, M. Küch, M. Stifter *et al.*, "Interfacing power system and ICT simulators: Challenges, state-of-the-art, and case studies," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 14–24, 2016, doi: 10.1109/TSG.2016.2542824.

[3] "IEEE standard for modeling and simulation (M&S) high level architecture (HLA)– framework and rules," *IEEE Std 1516-2010 (Revision of IEEE Std 1516-2000)*, pp. 1–38, Aug 2010, doi: 10.1109/IEEESTD.2010.5553440.

[4] "IEEE standard for modeling and simulation (M&S) high level architecture (HLA)– federate interface specification," *IEEE Std 1516.1-2010 (Revision of IEEE Std 1516.1-2000)*, pp. 1–378, Aug 2010, doi: 10.1109/IEEESTD.2010.5557728.

[5] M. Burns, T. Roth, E. Griffor, P. Boynton, J. Sztipanovits, and H. Neema, "Universal CPS environment for federation (UCEF)," in *2018 Winter Simulation Innovation Workshop*, 2018.

[6] (2019) Mak RTI. [Online]. Available: https://www.mak.com/

[7] (2018) Pitch pRTI. [Online]. Available: http://pitchtechnologies.com/

[8] G. Hemingway, H. Neema, H. Nine, J. Sztipanovits, and G. Karsai, "Rapid synthesis of high-level architecture-based heterogeneous simulation: a model-based integration approach," *Simulation*, vol. 88, no. 2, pp. 217–232, 2012, doi: 10.1177/0037549711401950.

[9] M. Maróti, T. Kecskés, R. Kereskényi, B. Broll, P. Völgyesi, L. Jurácz, T. Levendovszky, and Á. Lédeczi, “Next generation (meta) modeling: web-and cloud-based collaborative tool infrastructure,” *MPM@ MoDELS*, vol. 1237, pp. 41–60, 2014.

[10] H. Neema, J. Sztipanovits, C. Steinbrink, T. Raub, B. Cornelsen, and S. Lehnhoff, “Simulation integration platforms for cyber-physical systems,” in *Proceedings of the Workshop on Design Automation for CPS and IoT*. ACM, 2019, pp. 10–19, doi: 10.1145/3313151.3313169.

[11] M. Bombino and P. Scandurra, “A model-driven co-simulation environment for heterogeneous systems,” *International Journal on Software Tools for Technology Transfer*, vol. 15, no. 4, pp. 363–374, 2013, doi: 10.1007/s10009-012-0230-5.

[12] T. Nägele and J. Hooman, “Rapid construction of co-simulations of cyber-physical systems in HLA using a DSL,” in *2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 2017, pp. 247–251, doi: 10.1109/SEAA.2017.29.

[13] A. Falcone, A. Garro, S. J. Taylor, A. Anagnostou, N. R. Chaudhry, and O. Salah, “Experiences in simplifying distributed simulation: The HLA development kit framework,” *Journal of Simulation*, vol. 11, no. 3, pp. 208–227, 2017, doi: 10.1057/s41273-016-0039-4.

[14] P. T. Grogan and O. L. De Weck, “Infrastructure system simulation interoperability using the high-level architecture,” *IEEE Systems Journal*, vol. 12, no. 1, pp. 103–114, 2015, doi: 10.1109/JSYST.2015.2457433.

[15] (2014) Functional mock-up interface for model exchange and co-simulation 2.0. [Online]. Available: http://fmi-standard.org

[16] P. G. Larsen, J. Fitzgerald, J. Woodcock, P. Fritzson, J. Brauer, C. Kleijn, T. Lecomte, M. Pfeil, O. Green, S. Basagiannis *et al.*, “Integrated tool chain for model-based design of cyber-physical systems: The INTO-CPS project,” in *2016 2nd International Workshop on Modelling, Analysis, and Control of Complex CPS (CPS Data)*. IEEE, 2016, pp. 1–6, doi: 10.1109/CPSData.2016.7496424.

[17] E. Widl and W. Müller, “Generic FMI-compliant simulation tool coupling,” in *Proceedings of the 12th International Modelica Conference, Prague, Czech Republic, May 15-17, 2017*, no. 132. Linköping University Electronic Press, 2017, pp. 321–327.

[18] (2019) Portico RTI. [Online]. Available: http://www.porticoproject.org/

[19] T. Roth and M. Burns, “A gateway to easily integrate simulation platforms for co-simulation of cyber-physical systems,” in *2018 Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES)*. IEEE, 2018, pp. 1–6, doi: 10.1109/MSCPES.2018.8405394.

[20] T. Roth, M. Burns, and T. Pokorny, “Extending portico HLA to federations of federations with transport layer security,” *2018 Fall Simulation Innovation Workshop (SIW)*, no. 18F-SIW-038, 2018.

[21] B. Möller, S. Löf *et al.*, “Mixing service oriented and high level architectures in support of the GIG,” in *Proceedings of the 2005 Spring Simulation Interoperability Workshop*, no. 05S-SIW, 2005, p. 64.

[22] B. Möller and S. Löf, “A management overview of the HLA evolved web service API,” in *Proceedings of 2006 Fall Simulation Interoperability Workshop, 06F-SIW-024, Simulation Interoperability Standards Organization*, 2006.

Online Testbed for Evaluating Vulnerability of Deep Learning Based Power Grid Load Forecasters

Himanshu Neema, Peter Volgyesi, Xenofon Koutsoukos
Vanderbilt University
Nashville, TN 37235

Thomas Roth, Cuong Nguyen
National Institute of Standards and Technology
Gaithersburg, MD 20899

***Abstract*—Modern electric grids that integrate smart grid technologies require different approaches to grid operations. There has been a shift towards increased reliance on distributed sensors to monitor bidirectional power flows and machine learning based load forecasting methods (e.g., using deep learning). These methods are fairly accurate under normal circumstances, but become highly vulnerable to stealthy adversarial attacks that could be deployed on the load forecasters. This paper provides a novel model-based Testbed for Simulation-based Evaluation of Resilience (*TeSER*) that enables evaluating deep learning based load forecasters against stealthy adversarial attacks. The testbed leverages three existing technologies, viz. *DeepForge*: for designing neural networks and machine learning pipelines, *GridLAB-D*: for electric grid distribution system simulation, and *WebGME*: for creating web-based collaborative metamodeling environments. The testbed architecture is described, and a case study to demonstrate its capabilities for evaluating load forecasters is provided.**

***Index Terms*—power grid, load forecasting, machine learning, security, resilience, adversarial attacks, model-based testbed**

I. Introduction

This work has been motivated by our NSA Science of Security Lablet research efforts to create executable simulation models and repeatable experiments for evaluating potential vulnerabilities and successful resilient strategies for complex Cyber-Physical Systems. To address these needs we developed a web-based, cloud-hosted design environment and integrated state-of-the-art simulation engines for multiple CPS domains (highway and railway transportation, power distribution). This paper focuses on the power grid domain of our multi-model testbed.

In electrical grids, the power generation is typically conducted on-demand, which requires utilities to continuously forecast the grid loads [1]. The loads are estimated for the long-term (LT) (i.e., more than a year), medium-term (MT) (i.e., a month to a year), and short-term (ST) (i.e., an hour to a week). LT forecasts are used for planning the necessary generation, transmission, and distribution equipment. MT forecasts are used for adjusting the LT plans. ST forecasts are used for real-time grid operations and operating both the grid and power markets in a safe, secure, and reliable manner.

ST and MT forecasting have become challenging due to the dynamic and distributed nature of modern electrical grids. In the traditional grid, power is centrally generated, and the power flow is unidirectional from generation to transmission to distribution network. Smart grid technologies enable the integration of distributed energy resources (DERs) that provide local sources and introduce bidirectional power flow into the system. Additionally, most DERs are variable sources such as wind turbine and solar photovoltaic (PV) that are not always available due to changing weather conditions such as storms. Because of capability for bidirectional power flow and the variable nature of DER, grid operations need to evolve from a deterministic to a stochastic model.

Smart grid with DER integration have enabled consumers to generate power locally and provide excess generated power back to the electric grid for a financial benefit. In addition, the potential dynamic pricing of electricity necessitates the use of transactive controllers and smart demand management to get the best pricing of power (e.g., move consumption to off-peak hours). This stochastic environment has made load forecasting significantly challenging where traditional demand-supply and failure modeling is no longer suitable.

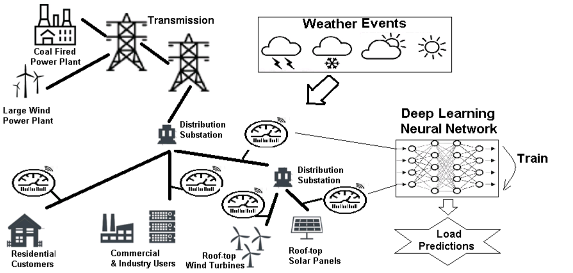


Figure 1: Deep Learning Based Load Forecasting in a Dynamic Power Grid with Distributed Energy Resources

To address the above challenges, smart grid deploys a large number of smart sensors that collect power flow readings at various points in the network. The measurements from these sensors can be used by deep learning based load forecasting systems to estimate the expected system loads. Figure 1 illustrates the variety of factors affecting ST load forecasting and how a deep learning system can be used to predict future loads. The sensor readings recorded by smart meters (generated by real system or a simulation) are stored in a time series database, which is used by a neural network to learn expected load on the system. Deep learning based forecasting is fairly accurate under normal circumstances, but, due to its complexity, becomes highly vulnerable to stealthy adversarial attacks. These

Neema, Himanshu; Volgyesi, Peter; Koutsoukos, Xenofon; Roth, Thomas; Nguyen, Cuong. "Online Testbed for Evaluating Vulnerability of Deep Learning Based Power Grid Load Forecasters." Paper presented at Modeling and Simulation of Cyber-Physical Energy Systems, Sydney, AU. April 21, 2020 - April 21, 2020.

attacks [2] subtly modify some of the sensor readings (by intercepting and forwarding modified data) such that not only the attacks remain undetectable by anomaly detection software, but also the accuracy of the load predictors is significantly reduced. As the adversarial attacks are not easily detected, they can result in cumulative disruptions leading to significant damage and loss before the system could be reverted to fallback methods.

Adversarial machine learning has been studied in many different categorization and estimation problems [2] [3]. Investigation of its use in cyber-physical systems (CPS) is recent, but is quickly emerging as an active research field [4]. The cyber-physical nature of smart grid makes it possible to attack specific cyber components such that the resulting failures cascade to a much wider region of the grid [5]. Further, attackers can exploit the growing networking capability of monitoring and control equipment to remotely attack specific components. Because the electric grid is one of the national critical infrastructures, it is crucial to study the vulnerability of different deep learning based forecasting methods earlier in the design cycle to avoid large-scale failures in a deployed system.

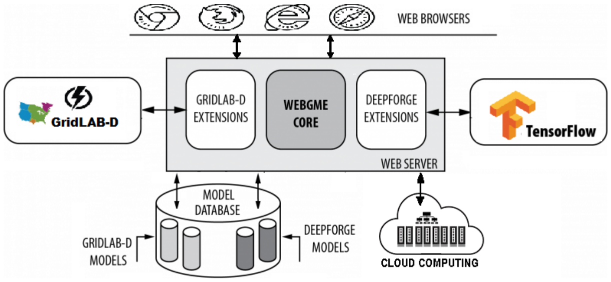


Figure 2: TeSER Testbed Architecture

In this paper, we describe a model-based Testbed for Simulation-based Evaluation of Resilience *(TeSER)* [6] that can analyze the resilience of load forecasters in the presence of stealthy adversarial attacks. As shown in Figure 2, the modeling front-end is built using the open-source Web-based Generic Modeling Environment (*WebGME*) [7] that supports creating web-based collaborative metamodeling environments. WebGME's plugin architecture enables interpretation of models and generation of system artifacts (e.g., source code, configuration files, and others), which can be used to execute experiments on a compute platform (e.g., desktop, server, or cloud). The stealthy adversarial attacks in this research were modeled using *DeepForge*, an open-source web-based environment for deep learning that enables collaborative modeling of neural networks and machine learning pipelines for reproducible deep learning experiments [8]. The testbed also utilizes an open-source power distribution system simulator, *GridLAB-D* [9], for simulating the electric grid and generating power flow data from smart sensors. In addition, the web-server contains an integrated *model database* that stores models of the electric grid, neural networks, and machine learning pipelines. TeSER executes the experiments on a connected *cloud computing* platform that enables the large-scale computations required by these experiments. The experiment results are collected and presented to the user as both raw data and digestible plots for analysis. The user is also provided with a full record of machine learning training iterations and console logs of the machine learning pipelines. We argue that by enabling earlier detection of the vulnerabilities in deep learning based load forecasting systems, our framework can help to both minimize the associated cost and make these systems more effective.

In the rest of the paper: Section II provides the motivation for analyzing vulnerabilities of load forecasters; Section III describes the core architectural components of TeSER; Section IV gives a case study to demonstrate TeSER's capabilities; and Section V concludes the paper and highlights future work. Note that a more detailed experiment—analyzing various attacker-defender games—is given in [10], this paper focuses on the architectural aspects of the testbed.

II. Motivation

Smart grid is a complex example of CPS [5] where the physical and computational components interact in specific ways to determine overall system functionality. For effective smart grid operations, a range of sensors are used to monitor aspects of the grid such as power flows at different locations, line continuity, equipment and device failures, actuator positions, and thermal characteristics. Load forecasting is essential for proper planning of the electrical system to determine the power equipment required and their arrangement, minimize system overloads, reduce power losses, manage operations effectively, and maintain the balance of supply and demand [1] [11].

Owing to bidirectional power flow and increases in DER integration, the smart grid has become highly dynamic, which directly affects the accuracy of load forecasts. The system dynamics are further affected by factors such as the variations in weather conditions and energy usage at different times of the day, price fluctuations in power markers, and a deeply integrated mix of residential, commercial, and industrial loads. Traditional methods of load forecasting use deterministic approaches (e.g., expert surveys and scenario-based assessment) and quantitative methods (e.g., time series analysis, smoothing averages and trend projections, least square estimates, and regression analysis). Since these methods tend to fit load expectations into a trending model, they do not work for highly dynamic variations in smart grid network topology (e.g., bidirectional power flow) and power supply and demand (e.g., DERs and time of use rate). Machine learning techniques can effectively handle these cases—where continuously updating the model is neither plausible nor pragmatic. The deep learning methods can be used for predicting loads in a more reliable manner while minimizing cost.

Deep learning methods, however, suffer from the *black box* problem, where there is no direct relation between inputs and outputs. Moreover, the electric grid has become more connected as the interactions between sensors, actuators, and controllers are largely enabled through a cyber communication network. This makes smart grids that use deep learning methods for load forecasting vulnerable to cyber-attacks and cascading failures. There are several examples where significant damage to the electric grid has occurred [12]. Therefore, the smart grid must be secured from potential cyber-attacks on deep learning based load forecasting methods. Stealthy adversarial attacks can intercept inputs of deep learning systems and subtly modify them to impact the system without being detected by anomaly detectors. These types of attacks on deep learning methods have been researched in the past, but as the deep learning methods are only recently being used in a CPS context, a testbed to enable such investigations is needed.

This paper describes a novel web-based and cloud-deployed testbed that can evaluate the vulnerability of deep learning models, that rely on a network of sensors for their inputs, to stealthy adversarial attacks. Specifically, the testbed is applied to load forecasters based on deep learning that use power flow readings from smart meters as their input. TeSER uses GridLAB-D for simulating an electric distribution system, and DeepForge for designing adversarial attack models and anomaly detectors. To avoid detection, the adversarial attacks are limited to modify the sensor readings within a given lower and upper threshold [13]. The testbed enables modeling of attacker-defender interactions as a *Stackelberg* game [4], where the defender uses a random subset of sensor readings and a neural network for load prediction, and then the attacker modifies some of those sensor readings within a configurable lower and upper bound. In addition, several load forecasters can be designed for introducing uncertainties that defend the load forecasting system against stealthy adversarial attacks.

III. TeSER Components & Features

The goal of TeSER is to provide a collaborative design and experimentation environment for evaluating the security and resilience of CPS amidst various cyber attack and defense strategies and the impact of these strategies on the physical infrastructure. The testbed leverages open-source technologies to build design tools that are reusable and configurable. It is web-based, cloud deployed, and supports real-time collaboration among researchers and analysts on models and experiments. In addition, it stores all input data, model parameters, and simulation results in the models, and version controls the models for experiment repeatability and provenance. This section provides an architectural overview of TeSER's core components and key features (as summarized in Figure 3).

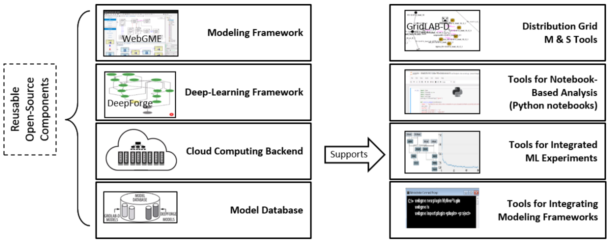


Figure 3: TeSER Core Components and Features

A. Modeling Framework

WebGME allows both the creation of rich, domain-specific modeling languages (DSMLs) and the use of those DSMLs to create domain models. It supports creating plugins that interpret the domain models, generate related system artifacts (e.g., source-code, scripts, configuration files), execute code on integrated compute platforms (e.g., cloud), collect experiment results, and display them to the users as digestible charts/plots and downloadable artifacts. For the modeling and experimentation front-end, WebGME's decorators and visualizers enable custom visualization of models, and the integrated console logging and feedback notifications allow users to get insights into long-running applications in the backend. A key aspect of WebGME is that it is web-based, highly scalable, and supports real-time collaboration among researchers and analysts who can edit the system and experiment models simultaneously from different locations using various web-browsers. WebGME stores all models in a MongoDB database and provides full version control and change tracking. Once designers create system models, analysts can use them to experiment with different designs. TeSER leverages WebGME for power distribution grid and deep learning models, but other CPS domains, such as transportation and healthcare, can also be supported using WebGME's extensible architecture.

B. Deep Learning Framework

DeepForge is built using WebGME and supports rapid development of neural-network and machine learning (ML) models. Modeling in DeepForge uses four main concepts, viz. *Operations*: atomic functions that accept named inputs and generate named outputs; *Pipelines*: specific ML activities such as training, data processing, and predicting; *Executions*: run-time instances of pipelines; and *Jobs*: run-time instances of operations along with associated execution metadata. As shown in Figure 4, the left side is a popular neural network model of a long-short term memory (LSTM) autoencoder for time series forecasting. The right side shows some of the reusable operations for creating ML pipelines (e.g., *GetLocalPredictor* implements a prediction routine and *PlotOperation* plots the time series data). A load forecasting processing pipeline is shown in the middle. Note, that the contents of some compound layers—most notably the LSTM blocks—are not modeled in DeepForge but directly mapped to classes in the Keras library.

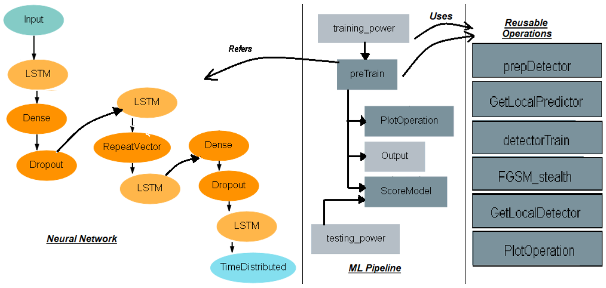


Figure 4: Deep Learning Framework

DeepForge allows users to add their own reusable operations in the library, and reuse pipeline models for creating variant or derived models. The neural network models are stored in a separate library for reuse. A neural network model can be used in multiple pipelines, and a single pipeline could use multiple neural network models. The integrated cloud backend is capable of executing several pipelines in parallel. DeepForge supports detailed views into pipeline execution progress by providing console logs that show the training iterations and by displaying integrated plots about training, testing, and prediction results. Another important feature of DeepForge is that it continuously aligns the Python code with corresponding pipeline models. The user can edit either the Python code or the pipeline model independently and DeepForge automatically synchronizes the other. In addition, all test data and models are stored in the connected artifacts store. With these comprehensive features and integrated experimentation support tools and compute infrastructure, DeepForge provides a powerful web-based environment for deep learning that TeSER leverages for creating its deep learning framework.

C. Distribution Grid Modeling & Simulation Tools

TeSER leverages a web-based platform from prior work [14] and WebGME for building its collaborative tools to support evaluation of distribution grids with integrated DER and their resilience against cyber and physical attacks. As shown in Figure 5, first a user either creates a new grid model or imports an existing GridLAB-D file and updates it as needed. Next, it provides *player* files (timestamped values of grid objects as input) and *recorder* files (specification of object values to collect as output). Finally, the *weather* files, if needed, can be supplied. The plugins are used to interpret the models and configurations to generate artifacts that are sent to a *Simulation Driver and Event Manager* module, which orchestrates the power grid simulation in the cloud accordingly, and gathers feedback from the executing experiments. The generated statistics specified in the recorder files are collected and returned to the user when the simulation completes. Continuous feedback is given while the simulation is running. TeSER uses this tool for simulating power distribution grid and generating smart meter recordings as input data for the deep learning models.

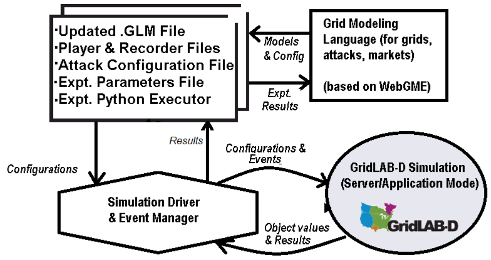


Figure 5: Distribution Grid Modeling & Simulation

D. Cloud Computing Backend

TeSER currently supports ML experiments in the power grid domain. The distribution grid simulation can take a long time for large grid models (e.g., with 1000+ nodes) or when multiple simulations are needed for different training inputs. The deep learning pipelines can also be computationally expensive, but multiple ML pipelines can be executed in parallel for faster responses. For these reasons, it is necessary to integrate a scalable and powerful compute platform. TeSER supports an integrated cloud computing backend that is hosted at Vanderbilt. The cloud also provides capability to store large datasets for ML pipelines and results of previous executions of pipelines. Users can login and inspect all of the executions ran previously (which could have taken hours or even days) including the console logs of all iterations, generated result files, and plots. Further, the user can investigate the execution results step-by-step and even re-execute any step (i.e., job).

E. Model Database

For web-based ML experimentation environments, it is crucial that models are version-controlled and accessible directly from the modeling frameworks. TeSER relies on the WebGME supported MongoDB object-oriented database. However, collaboration is highly challenging for large models (e.g., 1000+ node grid) as every small change in the interconnected graphical model can amount to a large update to be broadcasted. WebGME solves this issue by utilizing a Git-like commit architecture for model changes and only broadcasting deltas to all the modelers. The commits also enable fine-grained change tracking and allow relatively easier merging of models when a conflict arises. The WebGME model databse is used in TeSER to store not only the models of power grid, neural networks, machine learning pipelines, but also to store data and result artifacts. This greatly enhances the reproducibility and provenance of experiments. This database can also be queried using WebGME command-line tools, which enables automated experiments as well as testing.

F. Tools for Notebook-Based Analysis

CPS contain many interconnected physical and cyber components and their experimentation generates a large amount of data, which requires a rich framework for automated analysis. In the machine learning communities, Jupyter Notebooks are popular as they have integrated Python interpreter and analysis tools. TeSER has integrated Jupyter Notebooks in the WebGME front-end to facilitate analysis of data through Python scripts. As mentioned in Section III-B, DeepForge keeps the pipeline models and their corresponding Python code side-by-side and synchronized. The integration with Jupyter Notebooks allows embedding those Python scripts in Notebook cells and testing the corresponding pipeline models in an automated manner. For Notebook-based analysis, the user generates Jupyter Notebook files by executing TeSER plugins on a loaded model (or WebGME's Python libraries could be used to query the model from a user-created Notebook). TeSER also runs a Jupyter Notebook server besides the WebGME server and accesses it using a WebGME visualizer. The Notebook server enables users to develop scripts and algorithms for analyzing experiment results. As the Notebook server is accessed using WebGME visualizer, that can also update the model based on analysis results received from the Notebook server. Note that the notebook-based approach is optional for more advanced post-simulation analysis tasks.

G. Tools for Integrated ML Experiments

Experimentation with complex CPS not only requires parametric variations and cyber defense and attack combinations (i.e., design of experiments), but also an integrated deep-learning framework for analyzing ML algorithms. TeSER currently supports the power grid domain. It integrates GridLAB-D simulator for power distribution simulation and frameworks, such as Keras and TensorFlow, for creating deep-learning models. At present, the data from grid simulation is fed into deep learning models manually. However, we are working on integrating the two modeling environments into a single framework. This will support automated workflows of power grid simulation, generation of simulation data and its feed into corresponding ML pipelines. The pipeline execution results could also be fed back to update the grid model and close the loop. Figure 6 shows the overall architecture for integrated ML experiments. Here, the training operation in the ML pipeline refers to the corresponding grid model. The DeepForge code generator converts the pipeline model into an executable job script and the GridLAB-D Extension code generator generates templates to configure the simulation driver. We are developing an API for orchestrating the integrated ML experiments, which can execute grid simulations using an integrated simulation driver (like in III-C) and run the ML pipelines using the simulation data generated from grid simulations.

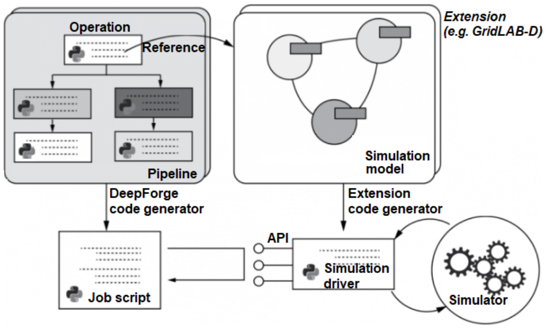


Figure 6: Integrated ML Experimentation in TeSER

H. Tools for Integrating Modeling Frameworks

Integrating modeling frameworks is needed in TeSER for connecting the power grid simulations and deep learning pipeline executions in a tight loop. We leverage WebGME's command-line interface for merging the corresponding DSML and creating an integrated modeling environment. In addition, we also leverage DeepForge's extension mechanism for integrating other modeling languages. The currently available support for creating Keras library based neural network models is an example of how DeepForge's extension mechanism works.

IV. Case Study

The collaborative model-based approach provided by TeSER enables rapid prototyping and experimentation with various neural network architectures and data processing, training and evaluation pipelines. Furthermore, tight integration with the CPS simulation tools—such as GridLAB-D—simplifies the process and shortens the time for input data generation for such models. Below we aim to illustrate how one can use the testbed to easily build and compare various ML-based predictors for load estimation in power distribution networks, and how this process is supported by the web-based design environment. For specific detailed analysis of prediction accuracy, its effects on load forecasting and of various attack and defend strategies, please refer to [10] [4].

A dataset has been generated with a realistic GridLAB-D model with 109 commercial and residential loads, where each endpoint reports its power usage on an hourly basis over a period of 90 days. Based on these reports the total power consumption in the distribution network is calculated. The case study aims to create and evaluate alternative regression models, which can predict the network-level (total) load for the next hour based on the previous 24 hours of data. The dataset was split into training and testing parts with 81 and 9 days of measurement data respectively.

We built a single generic data processing pipeline to train and evaluate the alternative ML models (see Figure 7). The testbed allows to (re)assign different ML architectures to the data processing pipeline. The current workflow loads the simulation results, implements the train/test split, drives the training (for a given number of epochs) and evaluates the model on the test data. It generates plot data for the training loss and summary statistics on the test results.

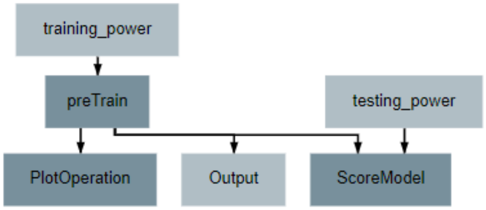
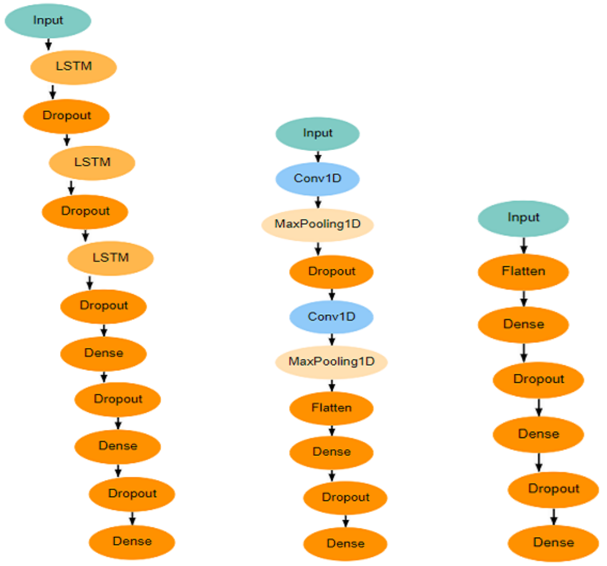


Figure 7: DeepForge generic pipeline model for training

As shown in Figure 8, we selected three popular architectures for time series forecasting with neural networks. First, a multilayer perceptron (MLP) model with two hidden layers (64-nodes each) is built. The second model is a deep convolutional network (CNN) with two sets of 1-dimensional filters sweeping on the time axis. Last, a more complex long short-term memory (LSTM) model is created (three LSTM layers with two fully connected layers). All three models use simple Rectified Linear Unit (*RelU*) activation functions and dropout layers for regularization. The training loss figure (Figure 9) of the three alternatives is captured from the testbed's web interface.



(a) LSTM model (b) CNN model (c) MLP model

Figure 8: Alternative neural network architectures for regression

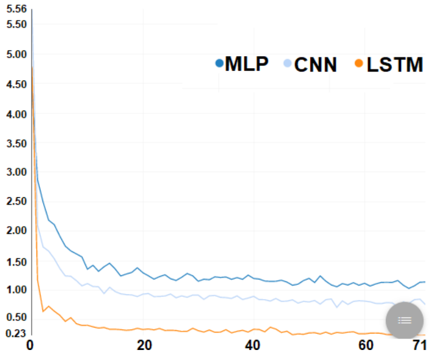


Figure 9: Training loss for MLP, CNN and LSTM

Evaluating the alternative predictors for the simulation data requires a single change in the pipeline–architecture assignment. The testbed takes care of input data generation, model versioning and tracing the analysis results back to the versioned model. This relatively simple experiment resulted in the prediction mean squared error (MSE) of 1.854 for MLP, 0.616 for CNN, and 0.106 for LSTM. As a baseline, the constant mean predictor has an MSE of 12.218. These results show the relative performances of the neural network models and agree with our assumptions on how 1-D convolutional and recurrent models can better learn and predict time series data.

V. Conclusions & Future Work

In this paper, we presented TeSER that is built using widely used open-source technologies. We presented its application in the power grid domain for evaluating deep learning based load forecasters. We described the testbed architecture and its core components and features and demonstrated it with a case study on load forecasters. TeSER uses a model-based approach and is web-based and cloud deployed. It provides strict version store for storing models (such as grid, neural network, and pipeline models), input datasets, and experiment results, thereby enabling experiments to be traceable and parameterizable.

We are working on integrating the power grid simulation and the deep learning framework into a single framework for end-to-end integrated workflows between them. This will support automated workflows of power grid simulation, generation of simulation data and its feed into corresponding ML pipelines. We are also extending our library of reusable and configurable cyber-attacks (modeled as ML pipelines), neural network models, and cyber-defense models of anomaly detectors that can mitigate the impact of stealthy adversarial attacks.

VI. Acknowledgments

This work is supported in part by the following grants: NSF #1521617, NIST #70NANB19H100, and NSA #H98230-18-D-0010. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor that such products are necessarily the best available for the purpose.

References

[1] Borges, Cruz E., et al., "Evaluating combined load forecasting in large power systems and smart grids," *IEEE Trans. on Industrial Informatics*, 2012, *doi*:10.1109/TII.2012.2219063.

[2] McDaniel, Patrick, et al., "Machine learning in adversarial settings," *IEEE Security & Privacy*, 2016, *doi*:10.1109/MSP.2016.51.

[3] Kurakin, Alexey, et al., "Adversarial machine learning at scale," 2016, *arXiv*:1611.01236v2.

[4] Ghafouri, Amin, et al., "Adversarial regression for detecting attacks in cyber-physical systems," 2018, *arXiv*:1804.11022v1.

[5] S. Soltan, M. Yannakakis, and G. Zussman, "React to cyber-physical attacks on power grids," *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 2, pp. 50–51, 2019, *doi*:10.1109/TNSE.2018.2837894.

[6] "Testbed for Simulation-based Evaluation of Resilience (TeSER)," February 2020, *URL*:http://lablet.webgme.org.

[7] Kecskés, Tamás, et al., "Bridging engineering and formal modeling: Webgme and formula integration." in *MODELS (Satellite Events)*, 2017, *Semanticscholar*:44617122.

[8] Broll, Brian, et al., "Deepforge: A scientific gateway for deep learning," *Gateways*, 2018, *doi*:10.6084/m9.figshare.7092272.v2.

[9] Chassin, David P., et al, "GridLAB-D: An open-source power systems modeling and simulation environment," in *IEEE PES T&D Conference and Exposition, 2008*, pp. 1–5, *doi*:10.1109/TDC.2008.4517260.

[10] Zhou, Xingyu, et al., "Evaluating Resilience of Grid Load Predictions under Stealthy Adversarial Attacks," *Resilience Week Symposium*, 2019, *doi*:10.1109/RWS47064.2019.8971816.

[11] N. Phuangpornpitak and W. Prommee, "A study of load demand forecasting models in electric power system operation and planning," *GMSARN Int. Journal*, 2016, *Semanticscholar*:52992311.

[12] H. He and J. Yan, "Cyber-physical attacks and defences in the smart grid: a survey," *IET Cyber-Physical Systems: Theory & Applications*, vol. 1, no. 1, pp. 13–27, 2016, *doi*:10.1049/iet-cps.2016.0019.

[13] Deka, Deepjyoti, et al., "Optimal data attacks on power grids: Leveraging detection & measurement jamming," in *2015 IEEE SmartGridComm*, pp. 392–397, *doi*:10.1109/SmartGridComm.2015.7436332.

[14] Neema, Himanshu, et al., "Web-Based Platform for Evaluation of Resilient and Transactive Smart-Grids," in *MSCPES 2019*. IEEE, 2019, pp. 1–6, *doi*:10.1109/MSCPES.2019.8738796.

On the Use of Machine Learning Models to Forecast Flashover Occurrence in a Compartment

Jun Wang[1], Wai Cheong Tam[1], Richard Peacock[1], Paul Reneke[1], Eugene Yujun Fu[2], Thomas Cleary[1], Grace Ngai[2], Hong-va Leong[2]

[1]*Fire Research Division, National Institute of Standards and Technology, Gaithersburg, MD, USA*
[2]*Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong*

Abstract

This study examines the potential use of machine learning algorithms to build a model to forecast the flashover occurrence in a single-floor, multi-room compartment fire. Synthetic temperature data for heat detectors in different rooms are generated and more than 1000 simulation cases are considered and a total of 8 million data points are utilized for model development.

The development of P-Flash (Prediction model for Flashover occurrence) is presented. Two special treatments, sequence segmentation and learning from fitting, are proposed to overcome the temperature limitation of heat detectors in real-life fire scenarios and to enhance prediction capabilities to forecast the flashover occurrence even with situations where there is no temperature data from all detectors. Experimental evaluation shows that P-Flash offers reliable prediction. The model performance is approximately 83 % and 81 %, respectively, for current and future flashover occurrence, considering heat detector failure at 150℃. Results demonstrate that P-Flash, a new data-driven model, may provide fire fighters real-time, trustworthy, and actionable information to enhance situational awareness, operational effectiveness, and safety for firefighting.

Keywords: Machine learning, flashover occurrence prediction, synthetic temperature data, detector failure, smart firefighting.

Introduction

Several research efforts have been conducted to develop data-driven models to predict flashover conditions in a compartment fire. Richards et al. [1] established an empirical relationship based on a wide range of test data. Provided the estimated heat release rate (HRR), fire location and compartment size could be obtained. Yet, the prediction is only suitable for one-room compartments. Based on a generic algorithm, Neviackas and Trouvé [2] obtained a generalized HRR which can be used to

Wang, Jun; Tam, Wai Cheong; Reneke, Paul A.; Peacock, Richard D.; Cleary, Thomas; Fu, Eugene Yujun; Ngai, Grace; Leong, Hong Va. "On the Use of Machine Learning Models to Forecast Flashover Occurrence in a Compartment." Paper presented at 17th International Conference on Automatic Fire Detection (AUBE 20) & Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2020), Mülheim an der Ruhr, DE. September 15, 2020 - September 17, 2020.

determine flashover conditions in multi-room geometries. Overholt and Ezekoye [3] also developed an inverse model using a predictor-corrected method to estimate the flashover occurrence. Based on smoke layer temperature measurements, the prediction accuracy of the model was shown to be within 60 s. However, a challenging problem exists in which all models [1-3] rely on a complete measurement dataset obtained from laboratory equipment. In practical situations, sensors such as heat and/or smoke detectors will stop functioning at a certain elevated temperature [4]. If the required temperature data is missing, the estimated HRR from these models will become highly uncertain and subsequently, the prediction of flashover based on the estimated HRR will be unreliable.

Unlike previous work [1-3], the temperature limitation for heat detectors is considered in this paper with the objective to develop a machine learning based model that can predict the current and the future flashover occurrence even with missing temperature data due to malfunctioning heat detectors.

Methodology

In this section, the development process of P-Flash (Prediction model for Flashover occurrence) is presented.

Data Generation

Consider a single-story building with three compartments as shown in Fig. 1. The dimensions of Room 1 are 3.5 m x 3.5 m and the dimensions of Room 2 and Corridor are 4.5 m x 4.5 m and 3.5 m x 1 m, respectively. The ceiling height is 2.5 m and it is identical for all compartments. For simplicity, the material of all walls, ceilings, and floors is gypsum wallboard. As seen in Fig. 1, there are 4 openings: 1) a window in Room 1, 2) a door between Room 1 and Corridor, 3) a door between Corridor and Room 2, and 4) an exit-door in Room 2. These openings are fully opened and their detailed geometric configurations are provided in [5]. There is one heat detector in each compartment and they are all located at the center of the compartment about 4.5 cm away from the ceiling. The outdoor conditions are typical with the temperature at 20 ˚C and atmospheric pressure of 101 kPa.

Given the geometric settings provided above, CFAST Fire Data Generator[1] (CData) [6] is used to execute 1000 simulation runs with a t-squared fire (which grows at a rate proportional to the time raised to the second power) at the center in Room 1 for a wide range of peak heat release rate (HRR) and time to HRR peak. Fig. 2 presents the mean temperature profiles as a function of time for three detectors. The dash

[1] CData is an in-house computer code that uses a validated fire simulation program, CFAST [7], as the simulation engine. One advantage of using CFAST is that synthetic data can be generated efficiently. In the future, more sophiscated fire simulation programs will be considered.

lines represent two times the standard deviation of detector temperature profiles over the 1000 different cases. It should be noted that more sophisticated building geometries and fires involving more complex stages (i.e. multi-floor buildings with ventilation controlled fires) can be considered in the future.

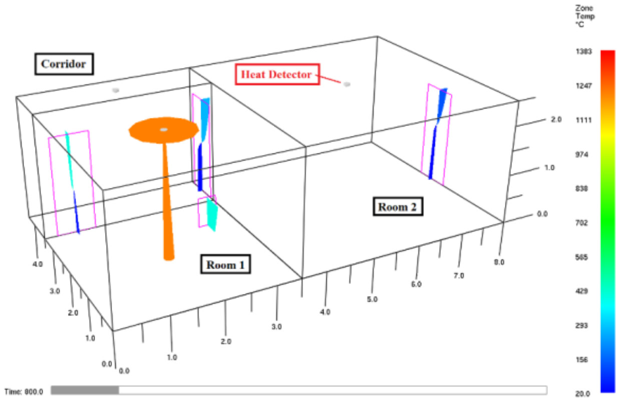


Fig 1. Schematic of the single-story multi-room compartment with a fire shown in the center of Room 1.

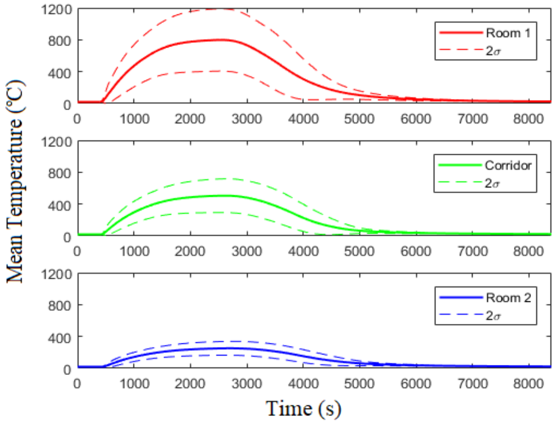


Fig. 2. Mean detector temperature profiles and two times its standard deviation in different compartments.

Sequence Segmentation for Detector Temperature Data

Loss of detector temperature signal is one of the major challenges for the development of a machine learning (ML) based flashover prediction model. For actual fire scenarios heat detectors cannot survive at elevated temperature [4] and would fail at temperatures well below flashover. In order to the overcome limited operating range, a special treatment to preprocess the detector temperature profiles is needed.

Assuming the detectors stop functioning at 150℃, Fig. 3 shows the detector ideal temperature profiles and those with a cut-off temperature at 150℃ for a simulation run with a fast-growth fire in Room 1. It can be seen that the available data for the Room 1 detector is very limited. In about 400 s, the temperature signal from Room 1 is lost and is artificially converted into a constant (i.e. a value of zero for simplicity). Yet, temperature signals from other compartments do exist. Given this observation, it is believed that the use of the available temperature data from other compartments will help to recover the detector temperature in Room 1 which can be used to determine the flashover condition from the room of fire origin.

Fig. 3 shows the implementation of sequence segmentation. The temperature profiles are divided into 4 phases. Each phase contains different available temperature signals. For example, signals from all detectors are present in Phase I. In Phase II, signals from the Corridor and Room 2 are available. In Phase III, only signals from Room 2 are left. Since it is well known that training a ML algorithm based on misleading data (i.e. zero temperature) jeopardizes the model accuracy, the use of sequence segmentation allows the ML algorithm to take advantage of the available signals from different phases. This treatment provides the basis for the model architecture of P-Flash.

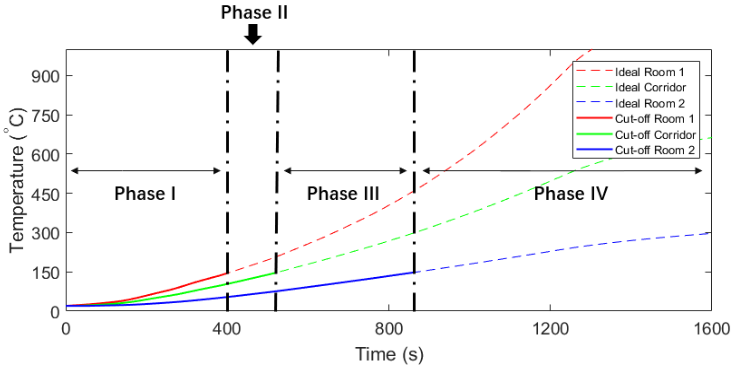


Fig. 3. Ideal detector and cut-off temperature for different phases.

P-Flash Architecture and Prediction Process

Fig. 4 depicts the model architecture and the prediction process of P-Flash. In general, P-flash consists of two regression models (R_{corr} and R_{R2}) and a memory component (M). Based on Fig. 3, the detector temperature in Room 1 is lost in Phase II and onward. For that, R_{corr} and R_{R2} are developed to predict/recover the temperature in Room 1 in different phases. The primary difference between the two models is that R_{corr} is trained based on the available information obtained from Corridor, and R_{R2} is trained based on Room 2. The memory component is a hybrid module: it performs as storage to contain outputs from R_{corr} and R_{R2} and to provide temperature prediction of Room 1 using the stored information. This architecture provides robust and flexible prediction capabilities to adapt to more complex cases with a larger number and different types of detectors. Since the prediction process of P-Flash requires a detailed explanation, the full description is provided in [5] due to the page limitations. However, key functions associated with procedures ① to ⑦ are highlighted below.

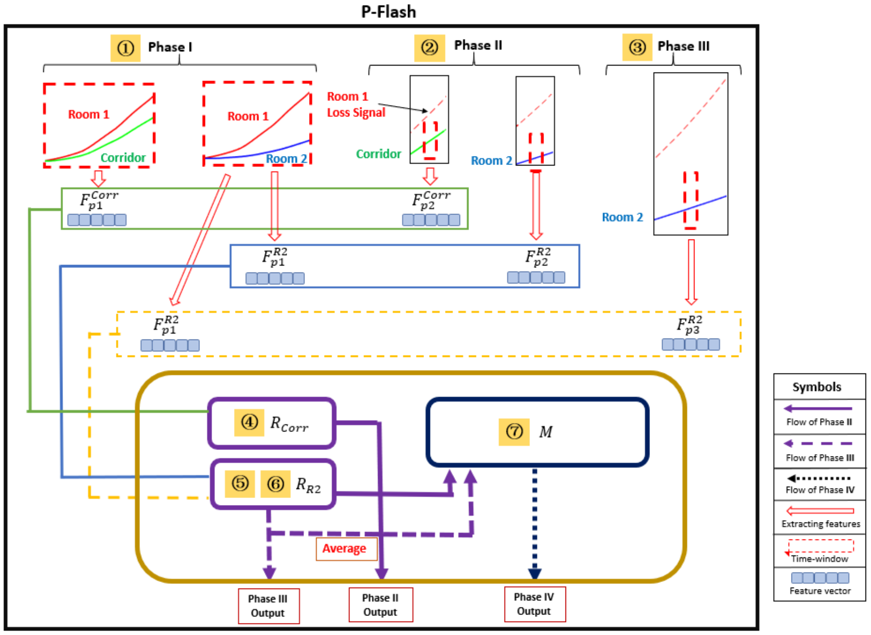


Fig. 4. Overview of P-Flash and the ML pipeline.

Feature Extraction (① to ③)

For development of the regression models, a set of features are needed. Typically, these properties are obtained based on feature extraction. As shown in Fig. 4, five different sets of features (*F*) are required to facilitate the model training process. It should be noted that one can use the original data as features (i.e. point-by-point temperature) to train a model. If the features are not extracted to relate data characteristics (i.e. rate of increase of temperature), it will be difficult for the model to learn any useful information (i.e. trends and/or patterns) for reliable predictions.

Given the sets of temperature data, feature vectors (F_{p1}^{Corr} and F_{p1}^{R2} in Phase I, F_{p2}^{Corr} and F_{p2}^{R2} in Phase II, and F_{p2}^{R2} in Phase III) are obtained based on delta and gradient[2]. Statistical features of delta and gradient are further extracted. In terms of notation, the superscript denotes the compartment that the features are used for training and the subscript denotes the phase from which the original data is being taken. For example, F_{p1}^{Corr} represents the feature vectors being used to train regression model, R_{corr}, based on temperature data obtained from Phase

[2] Delta refers to the overall temperature increase over the entire phase period and gradient refers to the rate of change of temperature increase through a time-window (see Fig. 4).

I. Readers can find the full list of features being obtained from Phase I, II, and III in [5].

Training and Testing (④ to ⑥)

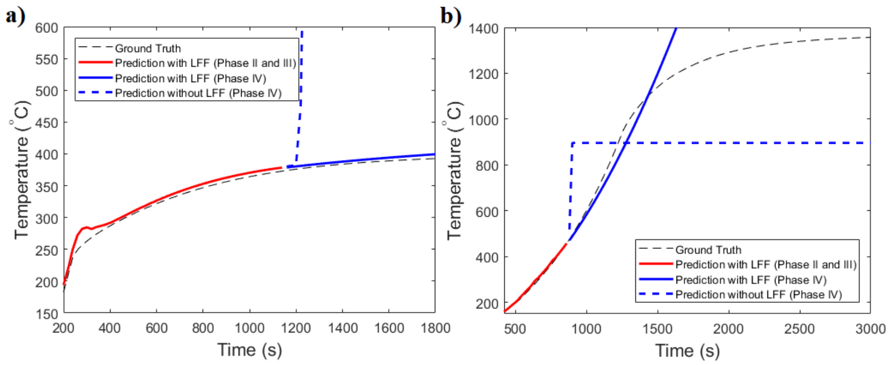
Support vector regression (SVR) [8] is used to develop the actual regression models (R_{corr} and R_{R2}). A 5-fold cross validation [8] method is applied to facilitate the training and testing process. In principle, the entire dataset from 1000 simulation runs is randomly divided into 5 subsets and each subset/fold contains 200 sessions. In general, one fold of data is being used as testing data and the remaining 4 folds are being used as training data. This process is carried out iteratively five times until all 5 different folds of data are being used as the testing set. The trained regression models provide Room 1 temperature predictions in Phase I to Phase III. Utilizing a grid search [8], the optimal configurations for SVR are C = 100 and Gamma = 0.05 with a radial basis function kernel.

Learning from Fitting (⑦)

In Phase IV, since all detectors are lost, no inputs are available and therefore no reliable predictions can be made from the regression models. In order to overcome this physical limitation, learning from fitting is implemented to facilitate the temperature prediction for Room 1 based on the historical temperature data of Room 1 (available temperature in Phase I and predicted temperature obtained in Phase II and III). As seen in Fig. 3, the increasing trend of the temperature in Room 1 is rather monotonic and follows similar behavior of a t-squared fire. Based on numerical experiment, a sigmoidal binding function is chosen for the fitting process. Mathematically, it is expressed as: $p_i = (b\sqrt{t_i})/(\sqrt{t_i} + a) + c$ where p_i is the prediction and t_i is the time associated with index *i*. Optimization is carried to obtain *a*, *b*, and *c* to produce a best fit to generalize the Room 1 temperature data in Phase I to III. Given the best fit, Room 1 temperature in Phase IV can be projected.

Results and Discussion

Figs. 5 show temperature predictions obtained from P-Flash for two selected cases: 1) a fast growth fire with low peak HRR case and 2) a medium growth fire with high peak HRR case. There are three sets of curves in each figure: i) ground truth, ii) prediction with learning from fitting (LFF), and iii) prediction without LFF. Each prediction curve is composed of up to two lines: a) the red line represents the Room 1 temperature predictions associated with Phase II and III and b) the blue line is for predictions in Phase IV. Since no prediction is needed for Phase I, the comparison is omitted. In Fig. 5a, P-Flash provides accurate temperature predictions of Room 1 in all phases and the benefit of using LFF is noticeable. After approximately 1150 s, when all detectors are lost, P-Flash is capable of providing predictions with similar trends and magnitudes. For P-Flash without LFF, the predictions rely on the regression models and it can be shown that the temperature prediction increases unrealistically to as high as 910℃. It is worth noting that the discrepancy observed at around 250 s is probably due to the change in temperature increase. Physically, it is the pivot point of its 2nd derivative where the rate of change of temperature increase changes from positive to negative. Additional work is under way to reduce such fluctuations.



Figs. 5. Comparison between ground truth and predictions obtained from P-Flash with and without LFF.

In Fig. 5b, it can be seen that the temperature of Room 1 being recovered from Phase II and III is in the fire growth stage. In the current version of P-Flash, it does not have additional information to predict the temperature decays. However, P-Flash is capable of predicting Room 1 temperature increase for flashover (i.e. temperature approaching 600℃). In order to evaluate the model performance over 1000 different cases, the mean absolute errors (MAE) are being calculated. Table 1 shows the MAE associated with different phases. It should be noted that the above results are denoted as "current prediction" and this prediction at time *t* is based on information obtained in time *t*.

Table 1. Performance summary for P-Flash with LFF.

	Phase II	**Phase III**	**Phase II & III**	**Phase IV**	**Overall Accuracy**
	MAE (℃)	**MAE (℃)**	**MAE (℃)**	**MAE (℃)**	**%**
Current Prediction	11.3	13.4	13.0	30.7	83.2
Future Prediction	12.9	15.5	13.4	37.6	81.7

In this work, it is important to evaluate the accuracy of the model in terms of classifying the occurrence of flashover. The overall accuracy is determined as the ratio of correct prediction within 20 s of the time of flashover to the total number of flashover occurrence in 1000 cases. As shown in Table 1, the model accuracy is approximately 83 % and 81 %, respectively, for the current and the future flashover occurrence.

On the fireground, knowing the room conditions of the fire origin is critical for optimizing rescue strategies and applying firefighting tactics. For that, P-Flash can forecast temperature in advance (i.e. 150 s) based on real-time information. As shown in Table 1, the MAE associated with future prediction only increases slightly with the same level of model accuracy.

Conclusion and Outlook

The development of P-Flash is presented. Based on experimental evaluation, P-Flash is capable of recovering required detector temperature for the determination of flashover in the room of fire origin. For practical use, P-Flash is under further development to handle more realistic and complicated cases. Our goal is that P-Flash will provide firefighters trustworthy and actionable information on the fireground to enable smart firefighting.

Acknowledgements

The authors wish to thank Dr. Michael Huang for helpful discussion.

References

[1] Richards, R., Munk, B., Plumb, O. (1997). Fire detection, location and heat release rate through inverse problem solution. Part I: theory. Fire Safety J 28(4):323–350.

[2] Neviackas, A., Trouvé, A. (2007). Sensor-driven inverse zone modeling of enclosure fire dynamics. In: SFPE Professional Development Conference and Exposition. Las Vegas, NV.

[3] Overholt, K. J., & Ezekoye, O. A. (2012). Characterizing heat release rates using an inverse fire modeling technique. Fire Technology, 48(4), 893-909.

[4] Pomeroy, A. T. (2010). Analysis of the effects of temperature and velocity on the response time index of heat detectors (Dissertation).

[5] Wang, J. & Tam, W. C. P-Flash – A Machine Learning based Flashover Prediction Model for Smart Firefighting (To be submitted).

[6] Peacock, R. D., Reneke, P. A. CFAST Fire Data Generator (CData): Monte Carlo Tools for CFAST (To be submitted).

[7] Peacock, R. D. & Reneke, P. A. & Forney, G. P. (2017). CFAST (Version 7): User's Guide. NIST Technical Note 1889v2.

[8] Yang, C.C. & Shieh, M.D. (2010). A support vector regression based prediction model of affective responses for product form design. Computers & Industrial Engineering, 59(4), 682-689.

Time Series Feature Extraction and Selection for Fire Data

Jun Wang[1], Youwei Jia[2], Eugene Yujun Fu[3], Jiajia Li[4], Wai Cheong Tam[1]
[1]Fire Research Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
[2]Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Guangdong, China
[3]Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong
[4]Department of Industrial Design, Guangdong University of Technology, Guangdong, China

Abstract

This paper aims to facilitate the use of machine learning to carry out supervised classification/regression tasks for time series data in fire research. Specifically, a feature engineering tool, FAST (Feature extrAction and Selection for Time-series), is developed. Using hypothesis test method together with principal component analysis, relevant features with high significance to the prediction are selected. A study case is presented for the use of FAST. Results demonstrate the importance of obtaining effective features and its potential benefits. It is expected that the feature engineering tool can help scientists and engineers in the fire research community develop accurate machine learning based models.

Keywords: Time series; detection; feature extraction; feature selection; machine learning

Introduction

Machine learning (ML) paradigm is among the top methods that can be used to overcome real-time detection in many disciplines. Cao and his co-worker [1] provided a support vector machine being trained based on historical financial data for unusual trading detection in stock markets. Using streaming signals from vibration sensors in multiple locations, Anton et al. [2] developed a neural network-based model to detect potential intrusion on an industrial machine. It was reported that the model can provide early warning to schedule maintenance thus avoiding mechanical failure for large scale, high cost equipment. Recently, research efforts are made to implement ML paradigms to overcome detection problems that traditional approaches such as physics-based models [3] might not be able to handle in the fire research community. Although the studies [4] showed that the use of ML paradigm provides

Wang, Jun; Jia, Youwei; Fu, Eugene Yujun; Li, Jiajia; Tam, Wai Cheong. "Time Series Feature Extraction and Selection Tool for Fire Data."
Paper presented at 17th International Conference on Automatic Fire Detection (AUBE 20) and Suppression, Detection and Signaling Research and Applications Conference (SUPDET 20), Mülheim an der Ruhr, DE. September 15, 2020 - September 17, 2020.

improvement in detection performance with significant reduction to nuisance alarms, two technical difficulties were identified.

Foremost, relevant data is limited. In fire research community, it can be easily shown that acquiring the desired data, such as temperature, heat flux, velocity, and species concentration, is not trivial because a) the event of interest (i.e. fire ignition in a compartment) does not happen frequently, b) time series data associated with fire event are not available to the public data warehouse [5], and c) physically conducting fire experiments is costly and time-consuming. Although fire simulation programs (i.e. Fire Dynamic Simulator [3]) can be used to generate synthetic data for ML task, the physical models being used in these codes are simplified and will not capable the correct physics of fundamental processes such as gasification and pyrolysis. For that, mutual collaboration between laboratories might have to be established to resolve the data problem.

In additional to problems associated with data availability, another technical difficulty for the use of ML paradigms in fire research community is that the data is complex. Figures 1 show two sets of time series data from the cooktop ignition study reported in [6]. Three selected measurements are shown: velocity, carbon monoxide (CO), and volatile organic compounds (VOCs). Since the measurements are made with different devices, each of the time series is being normalized to its peak value obtained for a test for illustration purpose. As it is shown in the figures, different profiles are observed for the VOCs data. While the data associated with oil has a monotonic increasing trend, the food data has multiple peaks. Also, data such as VOCs and CO (refer to left figure) have unphysical negative values with a number of missing data. More than that, the value associated with CO has a rapid increasing trend and becoming significant towards the end of a test. For velocity data, they fluctuate around a mean value. Lastly, it is worth noting that these measurements (time series) are rather non-stationary and this implies that the absolute values of these measurements are sensitive to the surroundings. These complexities make the application of the use of the ML paradigm difficult.

In order to train an accurate ML model, it is well understood that features[1] are required. Typically, feature engineering is being carried out to obtain effective features and this process requires an exhaustive amount of work. Although one can use the existing ML packages (i.e. tsfresh [7] and tslearn [8]), the obtained features are often irrelevant and difficult to interpret. Consequently, none of these ML packages are being used in fire research community. In this paper, a feature engineering tool, FAST

[1] Features can be considered as the transformed data that better represents the underlying problem to the predictive models which helps to improve the model accuracy on unseen data.

(Feature extrAction and Selection for Time-series), is introduced to facilitate the implementation of ML paradigm in fire research. In order to demonstrate the effectiveness of the use of FAST for development of a ML model, a case study is provided.

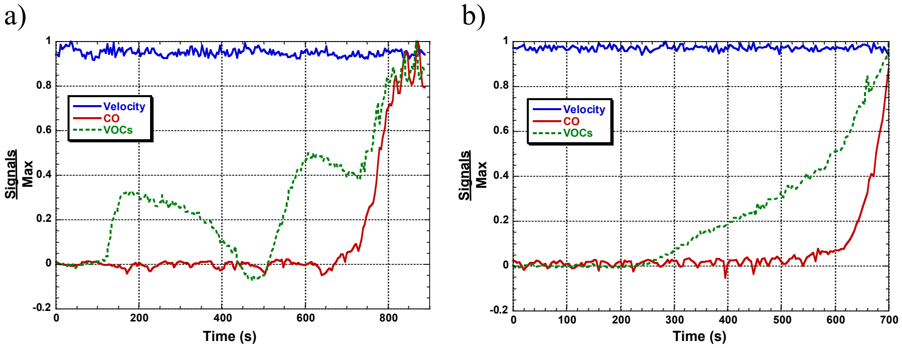


Figure 1. Time series data examples: a) measurements for a food and b) measurements for oil.

Development of FAST

The automated feature engineering tool, FAST, consists of two modules: 1) feature extraction and 2) feature selection. Figure 2 shows the overall structure of FAST and its operational procedures. In the following section,

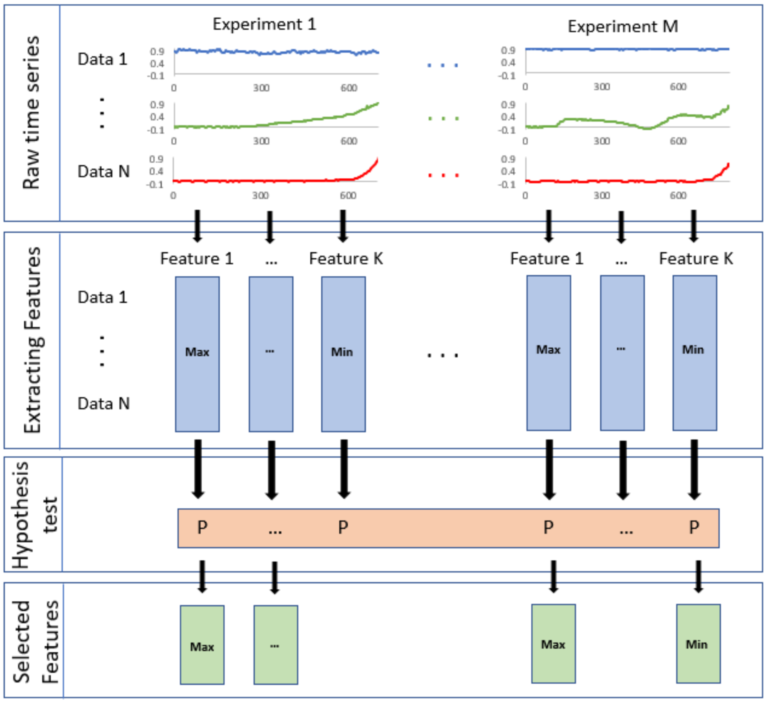


Figure 2. Overall structure of FAST and its operation procedures.

the details of each module and the descriptions of FAST operational procedures are presented. It should be noted that more effective features help to provide simpler models, better results, and higher numerical efficiency.

Feature extraction module (1 – 2)

In the current development of FAST, time series data, such as kitchen fire [6], smart firefighting technology [9], and wildland fire [10], are being considered. The behavior of these data varies significantly in which they can be either monotonic, periodic, or oscillatory. In order to capture the important tendency and/or pattern of these data, four types of feature extraction methods are being utilized and based on i) statistics, ii) trends, iii) its correlation, and iv) domain (i.e. frequency) of interest. In general, the application of statisticaly-based features such as mean or median over a sliding window[2] would help to obtain an average value and reduce the effect of data fluctuation due to outliers. For trend-based features, they provide information about the data with respect to time. In many regression problems, one might be interested in providing forecast of a future event and correlation-based features offer flexibility to obtain highly non-linear information. The last type of feature (frequency domain based) capture crucial information in frequency domain that is difficult to deduce in time domain. As shown in Fig. 2, the raw data will be separated into individual time series data in procedure 1. The data is then fed into the feature extraction module and a set of features (i.e. Min, Max, etc. as shown in procedure 2 in Fig. 2) are obtained. Table 1 provides a list of features that can be obtained in the feature extraction module. It should

Table 1. List of features being included in FAST.

Statistical based	Min	Max	Mean
	Median	Standard deviation	Sum
	Count above mean	Count below mean	Number of peaks
	Number of crossing		
Trend based	Delta	Rate of change	Acceleration
Correlation based	1st order regression	Max slope	Sinusoid
	2nd order regression	Exponential function	
Frequency domain based	Main frequency	Secondary frequency	
	Coefficients for discrete Fourier transform		

[2] A sliding window contains a specified length of time series data (i.e.30 seconds) and it moves over the data, sample by sample, and features are extracted over the data in the window.

be noted that since the module is written to perform the extraction procedure separately, additional features can be added if necessary.

Feature filtering module (3 – 4)

Two operations are carried out in the feature filtering module using hypothesis test and principal component analysis. The best 10 features are selected and they be readily used in training and testing of a ML algorithm.

Given the experimental dataset obtained in [7], which is partially being illustrated in Figs. 1, more than a hundred of features are obtained from the feature extraction module. The number of features would be 5 to 10 times more if methods from [9,10] are used. In general, having this large number of features is not ideal because the model can easily be overfitted and/or is hard to interpret. For that, a hypothesis test [12] is used to filter out the irrelevant features. In [7, 10, 11], the extracted features are usually continuous and labels can either be continuous and binary. For that, two test methods are adapted: i) Kolmogorov-Smirnov test [13] for cases where the features are continuous and the labels[3] are binary, and ii) Kendal rank test [14] for which both features and labels are continuous. Using either of the test methods, a p-value is obtained for each feature and the p-value quantifies the feature relevance to the prediction of the label. In principle, the smaller the p-value, the better/more relevant the features. Fig. 3 shows the top 10 selected features extracted from the kitchen fire dataset mentioned in [7] ranked by its p-value. In addition, two of the irrelevant features (most right) are also included in the figure. It can be seen that the velocity data is not useful and it can be understood based on the data behavior shown in Figs. 1.

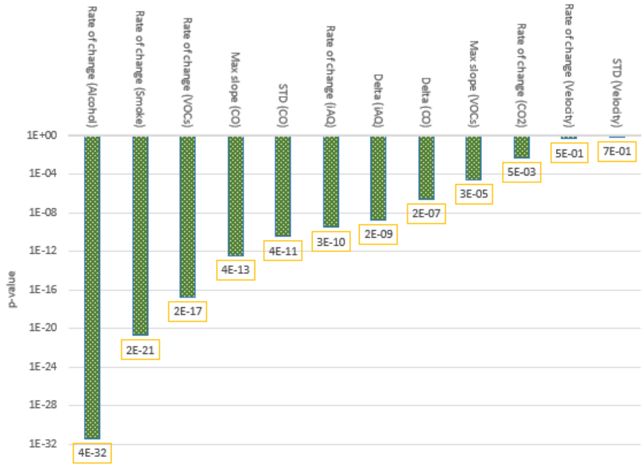


Figure 3. Selected features with small p-value.

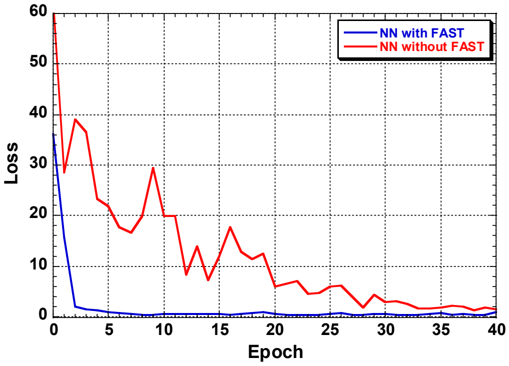
[3] Labels are referred to the targets for a classification/regression task.

Even though the hypothesis test helps to filter out a major portion of irrelevant features, there is still a high possibility that some of the selected features are redundant. For example, the features for median and mean are highly correlated in the absence of outliers in the time series. In order to avoid generating a group of highly correlated features, a principal component analysis (PCA) [15] is carried out and this step will help to select relevant features that are de-correlated.

Results and Discussion

In order to demonstrate the effectiveness of FAST, a case study is presented in this section. Given the set of time series obtained from [7], two sets of features are obtained: 1) one set with just the raw data and 2) one set using FAST. These two sets of features are used separately to facilitate the training of two individual 2-layer neural network (NN) models [15]. A holdout cross-validation [16] is used for training and testing. Approximately 60 % and 40 % of data is used as the training set and testing set, respectively. Similar to that of described in [7], the models attempt to classify abnormal cooking conditions based on the available sensor data.

For comparison, the two NN models are trained to have the same level of prediction accuracy[4]. Fig. 4 show the loss[5] as a function of epoch for two NN models. It can be seen that the model using FAST converges within 10 epochs and the model using raw data as features requires about four times more epochs to reach convergency. This highlights the benefits of obtaining relevant features for efficiency in training. Also, although both models are with 2-layer NN architecture, the NN model with FAST needs only 13 neurons for the first layer and 6 neurons for the second layer whereas the NN model without FAST requires 240 neurons in the first layer and 120 neurons in the second layer in order to achieve



Figs. 4. Loss for the two NN models during training.

[4] The accuracy is defined as the corrected predicted label over all possible labels.

[5] Loss value implies how well or poorly a certain model behaves after each iteration of optimization.

same level of prediction accuracy. It is worth noting that the prediction capability can generally be enhanced with the increasing number of layers and neuron. However, the model is highly possible to suffer from overfitting [17] in which the model is learning trends and patterns that are not used (i.e. the data oscillation as seen in Figs.1). Therefore, FAST can help model developers to avoid such training problems.

Conclusion and outlook

In this paper, the automated feature engineering tool, FAST (Feature extrAction and Selection for Time-series), is introduced. The overall structure of FAST and its operation procedure are described. Given a dataset similar to that of found in [7], 10 most-relevant features are obtained using FAST. The suggested features are then used in a case study to demonstrate its effectiveness. Results show that a machine learning model, such as a 2-layer neural network, trained using features obtained from FAST is more numerically efficient in terms of model training and has much simpler model structure. These observations highlight the importance of feature engineering. It is shown that the use of FAST will facilitate the development of an efficient and accurate machine learning model in fire research community.

Acknowledgement

The authors would like to thank Dr. Michael Huang and Dr. Wei Tang for the helpful discussion.

References

[1] Cao, L. J., & Tay, F. E. H. (2003). Support Vector machine with Adaptive Parameters in Financial Time Series Forecasting. IEEE Transactions on Neural Networks, 14(6), 1506-1518.

[2] Anton, S. D., Ahrens, L., Fraunholz, D., & Schotten, H. D. (2018, November). Time is of the Essence: Machine Learning-based Intrusion Detection in Industrial Time Series Data. In 2018 IEEE International Conference on Data Mining Workshops.

[3] Kevin McGrattan, K., Hostikka, S., McDermott, R., Floyd, J., & Vanella, M. (2018). Fire Dynamics Simulator User's Guide. NIST Special Publication, 1019, 1.

[4] Mensch, A., Hamins, A., Tam, W.C., Lu, J., Markell, K., You, C., & Kupferschmid, M. Sensors and Machine Learning Models to Prevent Cooktop Ignition and Ignore Normal Cooking. Submitted to Fire Technology.

[5] Asuncion, A., & Newman, D. (2007). UCI Machine Learning Repository.

[6] Mensch, A. E., Hamins, A. P., Lu, J., & Tam, W. C. (2019). Evaluating Sensor Algorithms to Prevent Kitchen Cooktop Ignition and Ignore Normal Cooking. SUPDET, Denvor, CO.

[7] Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests. *Neurocomputing*, *307*, 72-77.

[8] Tavenard, R. (2017). tslearn: A machine learning toolkit dedicated to time-series data. URL https://github. com/rtavenar/tslearn.

[9] Brown, C. U., Vogl, G. W., & Tam, W. C. (2019). Measuring Water Flow Rate for a Fire Hose Using a Wireless Sensor Network for Smart Fire Fighting. SUPDET, Denvor, CO.

[10] Stojanova, D., Panov, P., Kobler, A., Džeroski, S., & Taškova, K. (2006). Learning to Predict Forest Fires with Different Data Mining Techniques. In Conference on Data Mining and Data Warehouses, Ljubljana, Slovenia.

[11] Rice, J. A. (2006). Mathematical statistics and data analysis. Cengage Learning.

[12] Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association, 46, 68-78.

[13] Kendall, M. G. (1938). A new measure of rank correlation. Biometrika, 30(1/2), 81-93.

[14] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

[15] Ruck, D. W., Rogers, S. K., & Kabrisky, M. (1990). Feature selection using a multilayer perceptron. Journal of Neural Network Computing, 2(2), 40-48.

[16] Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

[17] Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference. Springer Verlag. New York.

Assessing Fire Smoke to Predict Backdraft and Smoke Explosion Potential

Ryan Falkenstein-Smith, Thomas Cleary
National Institute of Standards, Gaithersburg, USA

Abstract

The Fire Research Division at the National Institute of Standards and Technology is examining the ability to forecast backdraft or smoke explosions during a fire using a phi meter. Compared to other gas sensors, a phi meter can measure the global equivalence ratio under a broad spectrum of fuel-rich and fuel-lean conditions. Predetermined gas mixtures are fed into the phi meter wherein the equivalence ratio is measured. The measured and expected equivalence ratios for all gas mixtures are observed to be in substantial agreement. Concentration measurements are verified using the carbon to hydrogen ratio. Real-time temperature measurements are made at the inlet and outlet of the phi meter. The estimated enthalpy of the gas mixture from phi meter measurements is observed to increase as the gas mixture approaches stoichiometric conditions. The work demonstrates the potential to relate the estimated enthalpy in the phi meter with its oxygen consumption. The correlation can eventually be expanded to a variety of different fuels and conditions. Once established, the repository of data sets can be correlated to the bounding conditions for actual backdraft and smoke explosions of various strengths.

Keywords: Backdraft, Smoke Explosions, Phi Meter, Combustion Equivalence Ratio Meter, Flammability Limit, Gas Sampling

Introduction

The potential for backdrafts and smoke explosions to occur during fire events present a significant hazard to search and rescue teams, suppression teams, and other personnel in the surrounding area. Accurately assessing the condition of building spaces for potential smoke explosion or backdraft, before and during search and rescue and suppression activities, would improve situational awareness and lead to increased firefighter effectiveness and overall safety.

Falkenstein-Smith, Ryan; Cleary, Thomas. "Assessing Fire Smoke to Predict Backdraft and Smoke Explosion Potential." Paper presented at 17th International Conference on Automatic Fire Detection (AUBE 20) & Suppression, Detection and Signaling Research and Applications Conference (SUPDET 2020), Mülheim an der Ruhr, DE. September 15, 2020 - September 17, 2020.

A backdraft event is preceded by a change in the ventilation of a compartment filled with hot smoke that suddenly mixes with incoming Air through an opened door, broken window, or another venting source. Given the right conditions, the flammable gas mixture produced in the compartment ignites and expands, forcing fire gases out the opening, and subsequently erupting into a fireball or, with wind present, fire flower.

A smoke explosion is an event where a fuel-rich mixture of smoke (i.e., fuel vapors, carbon monoxide, and smoke particulate) located in an enclosed space mixes with oxygen (Air) that infiltrates the space. If a gas mixture is within the flammability limits while an ignition source is present, the mixture may explode with damaging deflagration overpressure. The fire service primarily relies on observations of smoke emanating from enclosures as evidence for the potential of backdraft or smoke explosions, such as puffing of the escaping smoke, color of the smoke, and other clues like darkened soot-covered window panes [1]. The fire service reliance on such indicators is not necessarily well-founded technically and is therefore not entirely reliable.

The full-scale backdraft experiments with No. 2 diesel fuel of Gottuk et al. [2] showed that the nominal fuel mass fraction (estimated from the fuel injection rate) correlated with the occurrence of backdraft and the resultant fireball size. The hypothesis proposed here is that the compartment gas temperature and the amount of fuel, oxygen, and non-combustible gases would allow the classification of the potential hazard. Furthermore, the potential hazard could be rated at varying strengths from no fireball or small overpressure, to large fireball extension or significant overpressure.

A significant hurdle to overcome in classifying backdraft is achieving gas measurements that accurately characterize the extracted sample. Extractive gas measurements with conventional catalytic bead flammable gas sensors are problematic for several reasons, some of which are: 1) the fuel composition is unknown which poses calibration issues; 2) the amount of particulate may play a role in these events and is not accounted for; and 3) these devices are typically limited to a range up to the lower flammability limit, they need sufficient oxygen to work, and poisoning or other loss of sensitivity is possible from the range of smoke exposures.

The idea proposed here is to measure the equivalence ratio, φ, using a phi meter operating under a new set of given conditions. The equivalence ratio is defined as the quotient of fuel to air ratio and the stoichiometric fuel to air ratio of the combustion process where fuel is initially un-combusted. Additional estimations and measurements include calculating the carbon to hydrogen ratio, an approximation of the fuel/oxygen ratio at the entrance of the phi meter, and the enthalpy increase from a combusted gas mixture. The measurements presented here lay the initial groundwork for developing a device that could predict potential backdrafts by quantifying certain conditions.

Experimental Method

In this work, the equivalence ratio is evaluated in real-time using a phi meter [3-7]. A phi meter has been constructed per the design of Babrauskas et al. [6] with the addition of thermocouples to measure the incoming and exiting flow temperatures. A flow diagram is provided in Fig.1. The phi meter operates by introducing extracted gas samples into a heated (900°C) plug flow reactor packed with platinum-coated beads as a combustion catalyst. An additional plug flow reactor placed within a furnace is available upstream to allow for the pre-combustion of fuel-air mixtures. Once the fuel/air mixture enters the phi meter, it is introduced to a counterflow of excess oxygen. Enough oxygen is supplied to the reactor such that the gas mixture is completely combusted.

Upon exiting the reactor, water is scrubbed from the gas stream using a chiller, resulting in a mixture primarily consisting of oxygen, carbon dioxide, and nitrogen. Assuming little to no nitrogen or oxygen in the fuel molecules, the known flow rates of the exit flow and added oxygen combined with a downstream measurement of the oxygen concentration allowed for a straightforward calculation of the equivalence ratio.

$$\varphi = \frac{{}^{Fuel}/_{Air}}{\left({}^{Fuel}/_{Air}\right)_{st}} = 1 + \left[\left(\frac{1 - X_{O_2,i}}{X_{O_2,i}\left(1 - X_{O_2,A} - X_{CO_2,A}\right)}\right)\left(\frac{n_{O_2}}{n_A} - X_{O_2,A}\right)\right] \quad \text{(Eq. 1)}$$

Here, n_{O_2} and n_A are the molar flow rates of oxygen introduced in the phi meter and total molar flow introduced into the analyzer, respectively. Additionally, $X_{O_2,i}$, $X_{O_2,A}$, and $X_{CO_2,A}$ are the volume fractions of oxygen in the Air (20.8%), and the dry oxygen and carbon dioxide introduced to the analyzer, respectively. A full derivation can be found in Ref. [6].

The upstream furnace is maintained at approximately 30°C, such that there is an un-combusted gas mixture entering the phi meter. The initial fuels of interest are methane and propane. The gas mixture varies within the flammability limits of the fuel [7] with φ_G ranging from approximately 0.5 to 2. The equivalence ratio is established from the predetermined fuel and air flows feeding into the upstream furnace. This work also has no additional air added to the flow stream past the upstream furnace (i.e., Air, 2=0, per Fig. 1). The fuel, Air, and oxygen volumetric flow rates are maintained using an Alicat Mass Flow Controller[1]. The volumetric flow rate exiting the chiller is measured using an Alicat Mass Flow Meter.

[1] Certain commercial products are identified in this report to specify adequately the equipment used. Such identification does not imply a recommendation by the National Institute of Standards and Technology, nor does it imply that this equipment is the best available for the purpose.

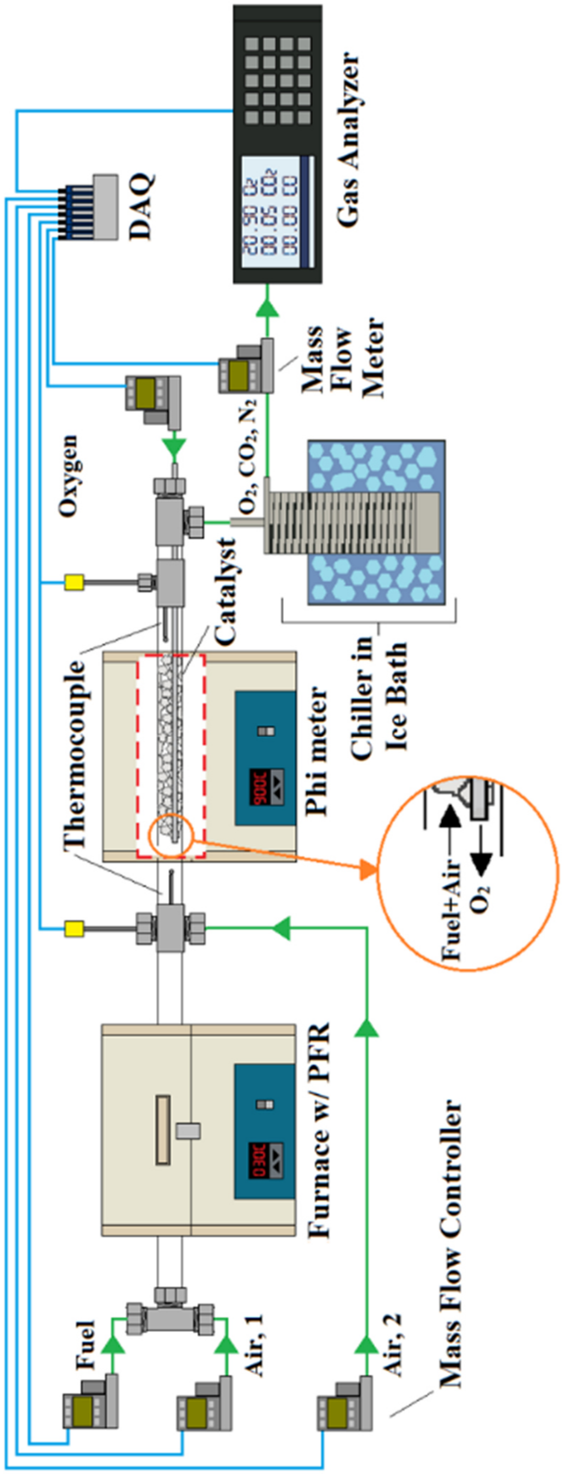


Fig. 1. A schematic of the phi meter experimental setup, including an additional furnace with a plug flow reactor (PFR) located upstream.

Oxygen and carbon dioxide concentrations are determined using a CAI 600 Series Gas Analyzer. Temperature measurements are made using two 20 gauge, 3 mm diameter, bare bead, Type K, low noise thermocouples positioned at the entrance and exit of the phi meter. The dataset is obtained approximately 10 min after the gas mixture composition entering the phi meter has been modified. All measurements are sampled at 2 Hz for approximately 2 min.

As a way to verify the accuracy of the result, the carbon to hydrogen mass ratio (C/H) is calculated for each analyzed sample using the equation below.

$$\frac{\mathrm{C}}{\mathrm{H}} = \frac{\mathrm{MW_C}}{\mathrm{MW_H}} \frac{\left(x_{\mathrm{CO_2}}\right) \frac{Y_{CO_2}}{\mathrm{MW}_{CO_2}}}{\left(y_{H_2O}\right) \frac{Y_{H_2O}}{\mathrm{MW}_{H_2O}}} \qquad \text{(Eq. 2)}$$

Here $\mathrm{x_{CO_2}}$ and $\mathrm{y_{H_2O}}$ are the number of carbon (1) and hydrogen (2) in carbon dioxide and water vapor, respectively. The molecular weight of carbon and hydrogen is denoted as $\mathrm{MW_C}$ and $\mathrm{MW_H}$, respectively. The mass fraction of species $\mathrm{CO_2}$, denoted as Y_{CO_2} is estimated from its measured volume fraction in the analyzer. The mass fraction of water vapor, Y_{H_2O}, is estimated from the difference between the total mass entering the phi meter and the total mass measured at the mass flow meter located past the chiller.

The uncertainties of all measurements are determined from a combined Type A and B evaluation of uncertainty. The Type A evaluation of uncertainty is estimated as the standard deviation of the measurement made during the sampling period. The Type B evaluation of uncertainty is estimated from the reported bias error in the instrumentation. The uncertainties presented in this work are expressed using a 95% confidence interval. For calculated values, the uncertainty was calculated using the law of propagation of uncertainty.

Results and Discussion

Fig. 2 presents a comparison between the expected and measured equivalence ratio. The expected equivalence ratio is calculated from the predetermined fuel and air flows feeding into the upstream furnace. The measured equivalence ratio is calculated using Eq. 1. As seen in Fig. 2, the measured values are equivalent, with some error, to their respective expected values for both fuel-rich and fuel-lean mixtures. The strong agreement indicates the validity of the phi meter set up, demonstrating a device that can measure the equivalence ratio of a fuel-air mixture.

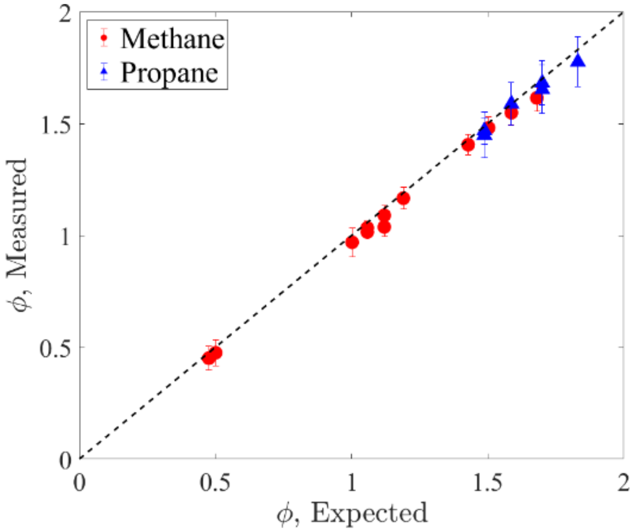


Fig. 2. A comparison between the expected and measured equivalence ratio. Note that the dotted line represents equivalency between the expected and measured values. The vertical bars are the estimated uncertainties of the measured equivalence ratio, calculated from the methods described in the previous section.

The limited propane data demonstrate a limitation of the phi meter design as conceived by the research engineer used in this work. The propane data presented here is limited by the flow capacity of the mass flow meter positioned downstream of the chiller. In this work, the maximum reading of the mass flow meter was 2 SLPM. In comparison to methane, propane requires a higher amount of Air to achieve stoichiometric combustion, which limited the investigation of propane to fuel-rich mixtures, given the mass flow meter constraint.

As previously stated, the carbon to hydrogen ratio is used to validate the accuracy of the measured volume fractions. As shown in Fig. 3, the estimated carbon to hydrogen ratio for each gas mixture is in strong agreement with the theoretical values of their respective fuels. This agreement validated the accuracy of the concentration measurements and the approximation of water concentration. These validation results suggest an additional feature of the phi meter; the calculated carbon to hydrogen ratio could be used to predict the composition of incoming fuels.

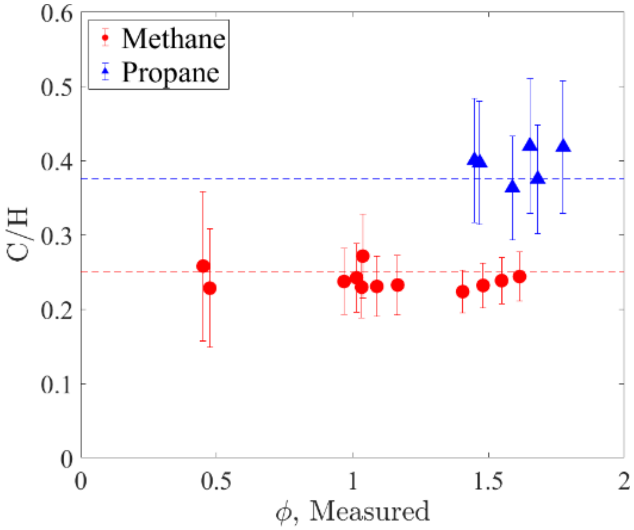


Fig. 3. Carbon to hydrogen ratio calculated from the product of all gas mixtures compared to the theoretical values at different equivalence ratios. The vertical bars on each data point represent the estimated uncertainty of the carbon to hydrogen ratio, calculated from the methods described in the previous section.

Fig. 4 presents the temperature difference across the phi meter for methane gas mixtures, maintained at a constant volumetric flow rate (2 SLPM), as a function of the total fuel to total oxygen ratio in the entire phi meter, (Fuel/O2) Tot. The temperature difference profile, as a function of the fuel/oxygen ratio, is not provided for propane since that obtained data is limited. The fuel/oxygen ratio is calculated from the un-combusted fuel and Air and the additional oxygen introduced to the mixture at the entrance of the phi meter. Since the objective of the phi meter is to completely combust the incoming fuel/air mixture, with the aid of additional oxygen, the combined gas mixture must always be lean such that $n_{\mathrm{Fuel}} < (n_{\mathrm{O_2}} + 0.21 \cdot n_{\mathrm{Air,1}})$. Based on that requirement, the fuel/oxygen ratio acts as an indicator of the operation of the phi meter. The fuel/oxygen ratio is calculated from the Eq.3, where $n_{\mathrm{Air,1}}$ is the molar flow rate of un-combusted Air flowing into the phi meter.

$$\left(\frac{\mathrm{Fuel}}{\mathrm{O_2}}\right)_{\mathrm{Tot.}} = \frac{n_{\mathrm{Fuel}}}{\left(n_{\mathrm{O_2}} + 0.21 \cdot n_{\mathrm{Air,1}}\right)} \qquad \text{(Eq. 3)}$$

The temperature difference across the phi meter as a function of the fuel/oxygen ratio is consistent for the methane fuel. The temperature is observed to increase as the fuel-lean gas mixture approaches stoichiometric conditions. For comparison to the calculated adiabatic flame temperatures, the temperature difference for each gas mixture is considerably less, indicating substantial heat loss within the reactor. Regardless, measuring the enthalpy increase of the reactor stream provides additional insight into the characteristics of the extracted gas mixture. The enthalpy measurement could be used to estimate the oxygen consumed internally based on the principle of oxygen consumption calorimetry.

The significance of the temperature measurement across the reactor is the additional information it can provide regarding the incoming fuel. As previously stated, the phi meter operation is based on the assumption that there is little to no oxygen and nitrogen in the fuel molecule. If the phi meter were to be used to investigate a fuel with a significant portion of oxygen or nitrogen in its composition, it could cause a significant error in the equivalence ratio measurement. The temperature measurements provide additional insight into the gas mixture introduced into the phi meter, which may prove useful when investigating unknown fuels.

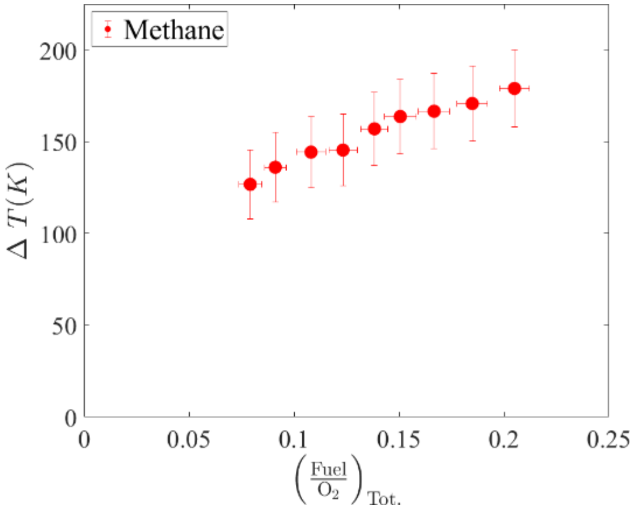


Fig. 4. The measured temperature difference across the phi meter as a function of the estimated fuel/oxygen ratio. It should be noted that the error of the temperature difference is determined from the measurement uncertainty derived from the uncertainty provided by the thermocouple manufacturer (±2% of the reading). The horizontal bars on each data point represent the uncertainty of the fuel to oxygen ratio within the phi meter and is calculated from the methods described in the previous section.

Conclusion and Future Work

This work highlights the application of a phi meter for measuring the equivalence ratio of methane and propane for fuel-rich and fuel-lean gas mixtures. This work demonstrates the phi meter's ability to measure the temperature rise across the reactor, which can be used if the future to infer the oxygen consumption within the phi meter and possibly provide additional information when dealing with fuels containing significant portions of oxygen and nitrogen. Future work will focus on expanding the library of fuels and gas mixtures to include gaseous fuels, with and without combustion gas diluents, and solid fuels like wood, plastics, and polyurethane foam combusted in low oxygen environments. Once an extensive library has been established, the phi meter will be applied to a 2/5 scale compartment, designed based on similar research [5,7,9], where backdraft and smoke explosions of various strengths can occur. The phi meter measurements can then be correlated to the bounding conditions for actual backdraft and smoke explosions of various strengths.

Acknowledgments

The authors would like to acknowledge Michael Selepak, who aided in the construction of the phi meter used for this work.

References

[1] International Fire Service Training Association, 1998. Essentials of Firefighting. Oklahoma State University.

[2] Gottuk, D.T., Williams, F.W., and Farley, J.P., 1997. The development and mitigation of backdrafts: a full-scale experimental study. Fire Safety Science, 5, pp.935-946.

[3] Andersson, B., Holmstedt, G., and Dagneryd, A., 2003. Determination of the equivalence ratio during fire, comparison of techniques. Fire Safety Science 7: 295-308.

[4] Blomqvist, P., and Lönnermark, A., 2001. Characterization of the combustion products in large-scale fire tests: comparison of three experimental configurations. Fire and Materials 25.2: 71-81.

[5] Parkes, A.R., 2009. The Impact of Size and Location of Pool Fires on Compartment Fire Behavior. University of Canterbury.

[6] Babrauskas, V., Parker, W.J., Mulholland, G. and Twilley, W.H., 1994. The phi meter: A simple, fuel-independent instrument for monitoring combustion equivalence ratio. Review of scientific instruments, 65(7), pp.2367-2375.

[7] Chen, N., 2012. Smoke Explosion in Severally Ventilation Limited Compartment Fires. University of Canterbury.

[8] Beyler, C., Flammability Limits of Premixed and Diffusion Flames, SFPE Handbook of Fire Protection Engineering M.J. Hurley (ed.) 2015. Springer.

[9] Fleischmann, C.M., Pagni, P.J., and Williamson, R.B., 1994. Quantitative Backdraft Experiments. Fire Safety Science 4: 337-348.

Cryptanalysis of LEDAcrypt

Daniel Apon[1], Ray Perlner[1], Angela Robinson[1], and Paolo Santini[2,3]

[1] National Institute of Standards and Technology, USA
[2] Università Politecnica delle Marche, Italy
[3] Florida Atlantic University, USA

{daniel.apon, ray.perlner, angela.robinson}@nist.gov
p.santini@pm.univpm.it

Abstract. We report on the concrete cryptanalysis of LEDAcrypt, a 2nd Round candidate in NIST's Post-Quantum Cryptography standardization process and one of 17 encryption schemes that remain as candidates for near-term standardization. LEDAcrypt consists of a public-key encryption scheme built from the McEliece paradigm and a key-encapsulation mechanism (KEM) built from the Niederreiter paradigm, both using a quasi-cyclic low-density parity-check (QC-LDPC) code. In this work, we identify a large class of extremely weak keys and provide an algorithm to recover them. For example, we demonstrate how to recover 1 in $2^{47.72}$ of LEDAcrypt's keys using only $2^{18.72}$ guesses at the 256-bit security level. This is a major, practical break of LEDAcrypt. Further, we demonstrate a continuum of progressively less weak keys (from extremely weak keys up to all keys) that can be recovered in substantially less work than previously known. This demonstrates that the imperfection of LEDAcrypt is fundamental to the system's design.

Keywords: NIST PQC, LEDAcrypt, McEliece, QC-LDPC, Cryptanalysis

1 Introduction

Since Shor's discovery [27] of a polynomial-time quantum algorithm for factoring integers and solving discrete logarithms, there has been a substantial amount of research on quantum computers. If large-scale quantum computers are ever built, they will be able to break many of the public-key cryptosystems currently in use. This would gravely undermine the integrity and confidentiality of our current communications infrastructure on the Internet and elsewhere.

In response, the National Institute of Standards and Technology (NIST) initiated a process [1] to solicit, evaluate, and standardize one or more quantum-resistant, public-key cryptographic algorithms. This process began in late 2017 with 69 submissions from around the world of post-quantum key-establishment mechanisms or KEMs (resp. public-key encryption schemes or PKEs), and digital signature algorithms. In early 2019, the list of candidates was cut from 69 to 26 (17 of which are PKEs or KEMs), and the 2nd Round of the competition began [2]. The conclusion of Round 2 is now rapidly approaching.

Apon, Daniel; Perlner, Ray; Robinson, Angela; Santini, Paulo. "Cryptanalysis of LEDAcrypt." Paper presented at Crypto 2020, Santa Barbara, CA, US. August 16, 2020 - August 20, 2020.

LEDAcrypt [16] is one of the 17 remaining candidates for standardization as a post-quantum PKE or KEM scheme. It is based on the seminal works of McEliece [20] in 1978 and Niederreiter [23] in 1986, which are based on the NP-complete problem of decoding an arbitrary linear binary code [5]. More precisely, LEDAcrypt is composed of a PKE scheme based on McEliece but instantiated with a particular type of codes (called QC-LDPC) and a KEM in the variant style of Niederreiter. The specific origins of LEDAcrypt – the idea of using QC-LDPC codes with the McEliece paradigm – dates back a dozen years to [15].

At a very high level, the private key of LEDAcrypt is a pair of binary matrices H and Q, where H is a sparse, quasi-cyclic, parity-check matrix of dimension $p \times p \cdot n_0$ for a given QC-LDPC code and where Q is a random, sparse, quasi-cyclic matrix of dimension $p \cdot n_0 \times p \cdot n_0$. Here p is a moderately large prime and n_0 is a small constant. The intermediate matrix $L = [L_0|...|L_{n_0-1}] = H \cdot Q$ is formed by matrix multiplication. The public key M is then constructed from L by multiplying each of the L_i by $L_{n_0-1}^{-1}$. Given this key pair, information can be encoded into codeword vectors, then perturbed by random error-vectors of a low Hamming weight.[1]

Security essentially relies on the assumption that it is difficult to recover the originally-encoded information from the perturbed codeword unless a party possesses the factorization of the public key as H and Q. To recover such matrices (or, equivalently, their product) one must find low-weight codewords in the public code (or in its dual) which, again, is a well-known NP-complete problem [5]. State-of-the-art algorithms to solve this problem are known as Information Set Decoding (ISD), and their expected computational complexity is indeed used as a design criteria for LEDAcrypt parameters.

The LEDAcrypt submission package in the 2nd Round of NIST's PQC process provides a careful description of the algorithm's history and specific design, a variety of concrete parameters sets tailored to NIST's security levels (claiming approximately 128-bit, 192-bit, and 256-bit security, under either IND-CPA or IND-CCA attacks), and a reference implementation in-code.

1.1 Our results

In this work, we provide a novel, concrete cryptanalysis of LEDAcrypt. Note that, in LEDAcrypt design procedure, the time complexity of ISD algorithms is derived by assuming that the searched codewords are uniformly distributed over the set of all n-uples of fixed weight. However, as we show in Section 3, for LEDAcrypt schemes this assumption does not hold, since it is possible to identify many families of secret keys, i.e., matrices H and Q, for which the rows of $L = HQ$ (which represent low weight codewords in the dual code) are characterized by a strong bias in the distribution of set bits. We define such keys as *weak* since, intuitively, in such a case an ISD algorithm can be strongly improved by taking into account the precise structure of the searched codeword. As a direct evidence, in Section 4 we consider a moderately-sized, very weak class

[1] We refer the reader to [3], A.1 for further technical details of the construction.

of keys, which can be recovered with substantially less computational effort than expected. This is a major, practical break of the LEDAcrypt cryptosystem, which is encapsulated in the following theorem.

Theorem 1.1 (Section 4). *There is an algorithm that costs the same as* $2^{49.22}$ *AES-256 operations and recovers* 1 *in* $2^{47.72}$ *of LEDAcrypt's Category 5 (i.e. claimed 256-bit-secure) ephemeral / IND-CPA keys.*

Similarly, there is an algorithm that costs the same as $2^{57.50}$ *AES-256 operations and recovers* 1 *in* $2^{51.59}$ *of LEDAcrypt's Category 5 (i.e. claimed 256-bit-secure) long-term / IND-CCA keys.*

While most key-recovery algorithms can exchange computational time spent vs. fraction of the key space recovered, this trade-off will generally be 1-to-1 against a secure cryptosystem. (In particular this trade off is 1-to-1 for the AES cryptosystem which is used to define the NIST security strength categories for LEDAcrypt's parameter sets.) However, we note in the above that both $49.22 + 47.72 = 96.94 \ll 256$ and $57.49 + 51.59 = 109.08 \ll 256$, making this attack quite significant. Additionally, we note that this class of very weak keys is present in every parameter set of LEDAcrypt.

While the existence of classes of imperfect keys is a serious concern, one might ask:

Is it possible to identify such keys during KeyGen, reject them, and thereby save the scheme's design?

We are able to answer this in the negative.

Indeed, as we demonstrate in Section 3, the bias in the distribution of set bits in L, which is at the basis of our attack, is intrinsic in the scheme's design. Our results clearly show that the existence of weaker-than-expected keys in LEDAcrypt is *fundamental* in the system's formulation and cannot be avoided without a major re-design of the cryptosystem.

Finally, we apply our new attack ideas to attempting key recovery without considering a weak key notion. Here we analyze the asymptotic complexity of attacking *all* LEDAcrypt keys.

Theorem 1.2 (Section 5). *The asymptotic complexity of ISD using an appropriate choice of structured information sets, when attacking* all *LEDAcrypt keys in the worst case, is* $\exp(\tilde{O}(p^{\frac{1}{4}}))$.

This gives a significant asymptotic speed-up over running ISD with uniformly random information sets, which costs $\exp(\tilde{O}(p^{\frac{1}{2}}))$. We note that simply enumerating all possible values of H and Q actually leads to an attack running in time $\exp(\tilde{O}(p^{\frac{1}{4}}))$, and indeed similar attacks were considered in LEDAcrypt's submission documents for the NIST PQC process. However, this type of attack had worse concrete complexity than ordinary ISD with uniformly random information sets for all of the 2nd Round parameter sets.

1.2 Technical Overview of Our New Attacks

Basic Approach: Exploiting the Product Structure. The typical approach to recovering keys for LEDAcrypt-like schemes is to use ordinary ISD algorithms, a class of techniques which can be used to search for low weight codewords in an arbitrary code. Generally speaking, these algorithms symbolically consider a row of an unknown binary matrix corresponding to the secret key of the scheme. From this row, they randomly choose a set of bit positions uniformly at random in the hope that these bits will (mostly) be zero. If the guess is correct and, additionally, the chosen set is an *information set* (i.e., a set in which all codewords differ at least in one position), then the key will be recovered with linear algebra computation. If (at least) one of the two requirements on the set is not met, then the procedure resets and guesses again.

For our attacks, intuitively, we will choose the information set in a non-uniform manner in order to increase the probability that the support of HQ, i.e. the non-zero coefficients of HQ, is (mostly) contained in the complement of the information set. At a high level, we will guess two sets of polynomials $H'_0, ..., H'_{n_0-1}$ and $Q'_{0,0}, ..., Q'_{n_0-1,n_0-1}$, then (interpreting the polynomials as $p \times p$ circulant matrices) group them into quasi-cyclic matrices H' and Q'. These matrices will be structured analogously to H and Q, but with non-negative coefficients defined over $\mathbb{Z}[x]/\langle x^p - 1\rangle$ rather than $\mathbb{F}_2[x]/\langle x^p + 1\rangle$. The hope is that the support of $H'Q'$ will (mostly) contain the support of HQ. It should be noted that a sufficient condition for this to be the case is that the support of H' contains the support of H and the support of Q' contains the support of Q. Assuming the Hamming weight of $H'Q'$ (interpreted as a coefficient vector) is chosen to be approximately W, then the information set can be chosen as the complement of the support of $H'Q'$ and properly passed to an ISD subroutine in place of a uniform guess.

Observe that the probability that the supports of H' and Q' contain the supports of H and Q, respectively, is maximized by making the Hamming weight of H' and Q' as large as possible while still limiting the Hamming weight of $H'Q'$ to W. An initial intuition is that this can be done by choosing the 1-coefficients of the polynomials $H'_0, ..., H'_{n_0-1}$ and $Q'_{0,0}, ..., Q'_{n_0-1,n_0-1}$ to be in a single, consecutive chunk. For example, by choosing the Hamming weight of the polynomials (before multiplication) as some value $B \ll W$, we can take $H'_0 = x^a + x^{a+1} + ... + x^{a+B-1}$ and $Q'_{0,0} = x^c + x^{c+1} + ... + x^{c+B-1}$.

Note that the polynomials H'_0 and $Q'_{0,0}$ (chosen with consecutive 1-coefficients as above) have Hamming weight B, while their product only has Hamming weight $2B - 1$. In the most general case, uniformly chosen polynomials with Hamming weight B would be expected to have a product with Hamming weight much closer to $\min(B^2, p)$. That is, for a fixed weight W required of $H'Q'$ by the ISD subroutine, we can guess around $W/2$ positions at once in H' and Q' respectively instead of something closer to $\sqrt{W}$ as would be given by a truly uniform choice of information set. As a result, each individual guess of H' and Q' that's "close" to this outline of our intuition will be more rewarding for searching the keyspace than the "typical" case of uniformly guessing information sets.

This constitutes the core intuition for our attacks against LEDAcrypt, but additional considerations are required in order to make the attacks practically effective (particularly when concrete parameters are considered). We enumerate a few of these observations next.

Different ring representations. The idea of choosing the polynomials within H' and Q' with consecutive nonzero coefficients makes each iteration of an information set decoding algorithm using such an H' and Q' much more effective than an iteration with a random information set. However there is only a limited number of successful information sets with this form. We can vastly increase our range of options by observing that the ring $\mathbb{F}_2[x]/\langle x^p+1\rangle$ has $p-1$ isomorphic representations which can be mapped to one another by the isomoprhism $f(x) \to f(x^\alpha)$. This allows us many more equally efficient choices of the information set, since rather than restricting our choices to have polynomials H'_0 and $Q'_{0,0}$ with consecutive ones in the standard ring representation, we have the freedom to choose them with consecutive ones in any ring representation (provided the same representation is used for H'_0 and $Q'_{0,0}$.)

Equivalent keys. For each public key of LEDAcrypt, there exist many choices of private keys that produce the same public key. In particular, the same public key $M = (L_{n_0-1})^{-1}L$ produced by the private key

$$H = [H_0, H_1, \cdots, H_{n_0-1}],$$

$$Q = \begin{bmatrix} Q_{0,0} & Q_{0,1} & \cdots & Q_{0,n_0-1} \\ Q_{1,0} & Q_{1,1} & \cdots & Q_{1,n_0-1} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{n_0-1,0} & Q_{n_0-1,1} & \cdots & Q_{n_0-1,n_0-1} \end{bmatrix};$$

would also be produced by any private key of the form

$$H' = [x^{a_0}H_0, x^{a_1}H_1, \cdots, x^{a_{n_0-1}}H_{n_0-1}],$$

$$Q' = \begin{bmatrix} x^{b-a_0}Q_{0,0} & x^{b-a_0}Q_{0,1} & \cdots & x^{b-a_0}Q_{0,n_0-1} \\ x^{b-a_1}Q_{1,0} & x^{b-a_1}Q_{1,1} & \cdots & x^{b-a_1}Q_{1,n_0-1} \\ \vdots & \vdots & \ddots & \vdots \\ x^{b-a_{n_0}}Q_{n_0-1,0} & x^{b-a_{n_0}}Q_{n_0-1,1} & \cdots & x^{b-a_{n_0}}Q_{n_0-1,n_0-1} \end{bmatrix};$$

for any integers $0 < a_i, b < p, i \in \{0, \ldots, n_0-1\}$. These p^{n_0+1} equivalent keys improve the success probability of key recovery attacks as detailed in the following sections.

Different degree constraints for H' and Q'. While we have so far described H' and Q' as having the same Hamming weight B, this does not necessarily need to be the case. In fact, there are many, equivalent choices of H' and Q' which

Apon, Daniel; Perlner, Ray; Robinson, Angela; Santini, Paulo. "Cryptanalysis of LEDAcrypt." Paper presented at Crypto 2020, Santa Barbara, CA, US. August 16, 2020 - August 20, 2020.

produce the same product $H'Q'$ based on this observation. For example, the product of

$$H'_0 = x^a + x^{a+1} + ... + x^{a+B-1}$$
$$Q'_{0,0} = x^c + x^{c+1} + ... + x^{c+B-1}$$

is identical to the product of

$$H'_0 = x^a + x^{a+1} + ... + x^{a+B-1-\delta}$$
$$Q'_{0,0} = x^c + x^{c+1} + ... + x^{c+B-1+\delta}$$

for any integer $-B < \delta < B$. More generally, this relationship (that if H' shrinks and Q' proportionally grows, or vice versa, then the product $H'Q'$ is the same) is independently true for any set of $\{H'_i, Q'_{i,0}, ..., Q'_{i,n_0-1}\}$ for $i \in \{0, ..., n_0 - 1\}$.

Attacks for $n_0 = 2$ imply similar-cost attacks for $n_0 > 2$. Our attacks are more easily described (and more effective) in the case $n_0 = 2$. In this case, we apply ISD to find low-weight codewords in the row space of the public key $[M_0 \mid M_1]$ to recover a viable secret key for the system. Naively extending this approach for the case $n_0 > 2$ to the entire public key $[M_0 \mid ... \mid M_{n_0}]$ requires constraints on the support of $n_0 + n_0^2$ polynomials (n_0 polynomials corresponding to H' and n_0^2 polynomials corresponding to Q'), so the overall work in the attack would increase quadratically as n_0 grows. However, even in the case that $n_0 > 2$, we observe that it is sufficient to find low weight codewords in the row space of only $[M_0 \mid M_1]$ in order to recover a working key, implying that the attack only needs to consider $3n_0$ polynomials $H_i, Q_{j,0}, Q_{k,1}$. So, increasing n_0 will make all of our attacks less effective, but not substantially so. More importantly, any attack against $n_0 = 2$ parameters immediately implies a similar-cost attack against parameters with $n_0 > 2$. Therefore, we focus on the case of $n_0 = 2$ in the remainder of this work.

A Continuum of Progressively Less Weak Keys. The attacker can recover keys with the highest probability per iteration of ISD by using a very structured pattern for L'. As we will see in Section 4, in this pattern both L'_0 and L'_1 will have a single contiguous stretch of nonzero coefficients in some ring representation. The result is a practical attack, but one which is only capable of recovering weak keys representing something like 1 in 2^{40} or 1 in 2^{50} private keys.

However, if the attacker is willing to use a more complicated pattern for the information set, using different ring representations for different blocks of H' and Q', and possibly having multiple separate stretches of consecutive nonzero coefficients in each block, then the attacker will not recover keys with as high a probability per iteration, but the attack will extend to a broader class of slightly less weak keys. This may for example lead to a somewhat less practical attack that recovers 1 in 2^{30} keys, but still much faster than would be expected given the claimed security strength of the parameter set in question.

We do not analyze the multitude of possible cases here, but we show they must necessarily exist in Section 3 by demonstrating that bias is intrinsically present throughout the LEDAcrypt key space.

Improvements to Average-case Key Recovery. In Section 5 we will take the continuum of progressively weaker keys to its logical extreme. We show that the attacks in this paper are asymptotically stronger than the standard attacks not just for weak keys, but for all keys.

As we move away from the simpler information set patterns used on the weakest keys, the analysis becomes more difficult. To fully quantify the impact of our attack on average keys would require extensive case analysis of all scenarios that might lead to a successful key recovery given a particular distribution of information sets used by the attacker, which we leave for future work.

1.3 Related Work

The main attack strategies against cryptosystems based on QC-LDPC codes are known as information set decoding (ISD) algorithms. These algorithms are also applicable to a variety of other code-based cryptosystems including the NIST 2nd round candidates BIKE [22], HQC [8], Classic McEliece [9], and NTS-KEM [17]. Initiated by Prange [25] in 1962, these algorithms have since experienced substantial improvements during the years [4, 7, 12, 13, 18, 19, 28]. ISD algorithms can also be used to find low-weight codewords in a given, arbitrary code. ISD main approach is that of guessing a set of positions where such codewords contain a very low number of set symbols; when this set is actually an information set, then linear algebra computations yield the searched codeword (see [3], Appendix A.3). ISD time complexity is estimated as the product between the expected number of required information set guesses and the cost of testing each set. Advanced ISD algorithms improve Prange's basic idea by reducing the average number of required guesses, at the cost of increasing the time complexity of the testing phase. Quantum ISD algorithms take into account Grover's algorithm [10] to quadratically accelerate the guessing phase. A quantum version of Prange's algorithm [6] was presented in 2010, while quantum versions of more advanced ISD algorithms were presented in 2017 [11].

In the case of QC-MDPC and QC-LDPC codes, ISD key recovery attacks can get a speed-up which is polynomial in the size of the circulant blocks [26]. This gain is due to the fact that there are more than one sparse vector in the row space of the parity check matrix, and no modification to the standard ISD algorithms is required to obtain this speed-up. Another example of gains due to the QC structure is that of [14] which, however, works only in the case of the circulant size having a power of 2 among its factors (which is not the case we consider here).

ISD can generally be described as a technique for finding low Hamming-weight codewords in a linear code. Most ISD algorithms are designed to assume that the low-weight codewords are random aside from their sparsity. However, in some cryptosystems that can be cryptanalyzed using ISD, these short codewords are not random in this respect, and modified versions of ISD have been used to break these schemes [21, 24]. Our paper can be seen as a continuation of this line of work, since unlike the other 2nd Round NIST candidates where ISD is cryptanalytically relevant, the sparse codewords which lead to a key recovery of

LEDAcrypt are not simply random sparse vectors, but have additional structure due to the product structure of LEDAcrypt's private key.

2 Preliminaries

2.1 Notation

Throughout this work, we denote the finite field with 2 elements by $\mathbb{F}_2$. We denote the Hamming weight of a vector a (or a polynomial a, viewed in terms of its coefficient vector) as $\mathrm{wt}(a)$. For a polynomial a we use the representation $a = \sum_{i=0}^{p-1} a_i x^i$, and call a_i its i-th coefficient. We denote the support – i.e. the non-zero coordinates – of a vector (or polynomial) a by $\mathrm{S}(a)$. In similar way, we define the *antisupport* of a, and denote it as $\bar{\mathrm{S}}(a)$, as the set of positions i such that $a_i = 0$. Given a polynomial a and a set J, we denote as $a|_J$ the set of coefficients of a that are indexed by J. Given π, a permutation of $\{0, \cdots, n-1\}$, we represent it as the ordered set of integers $\{\ell_0, \cdots, \ell_{n-1}\}$, such that π places ℓ_i in position i. For a length-n vector a, $\pi(a)$ denotes the action of π on a, i.e., the vector whose i-th entry is a_{ℓ_i}. For a probability distribution $\mathcal{D}$, we write $X \sim \mathcal{D}$ if X is distributed according to $\mathcal{D}$.

2.2 Parameters

The parameter sets of LEDAcrypt that we explicitly consider in this work are shown in Table 1 (although similar forms of our results hold for all parameter sets). We refer the reader to [3], Appendix A.1 for further technical details of the construction.

NIST Category	Security Type	p	d_v	m_0	m_1	n_0
1 (128-bit)	IND-CPA	14,939	11	4	3	2
5 (256-bit)	IND-CPA	36,877	11	7	6	2
5 (256-bit)	IND-CCA	152,267	13	7	6	2

Table 1. LEDAcrypt parameter sets that we consider in this paper.

3 Existence of Weak Keys in LEDAcrypt

As we have explained in Section 1.3, key recovery attacks against cryptosystems based on codes with sparse parity-check matrices can be performed by searching for low weight codewords, either in the code or in its dual. For instance, such codewords in the dual correspond, with overwhelming probability, to the rows of the secret parity-check matrix, of weight $\omega \ll n$, where n denotes the code length. The most efficient way to solve this problem is to use ISD algorithms. To analyze the efficiency of such attacks, weight-ω codewords are normally modeled as independent random variables, sampled according to the uniform distribution of n-uples with weight ω, which we denote as $\mathcal{U}_\omega$. At each ISD iteration, the algorithm succeeds if the intersection between the chosen set T and the support of (at least) one of such codewords satisfies some properties. Regardless of the considered ISD variant, this intersection has to be small.

Let ϵ be the probability that a single ISD iteration can actually recover a specific codeword of the desired weight. When the code contains M codewords of weight ω, then the probability that a single ISD iteration can recover any of these codewords is $1-(1-\epsilon)^M$ which, if $\epsilon M \ll 1$, can be approximated as ϵM. This speed-up in ISD algorithm normally applies to the case of QC codes, where M corresponds to the number of rows in the parity-check matrix (that is, $M = n-k$).

In this section we show that the product structure in LEDAcrypt yields to a strong bias in the distribution of set symbols in the rows of the secret parity-check matrix $L = HQ$. As a consequence, the assumption on the uniform distribution of the searched codewords does not hold anymore, and this opens up for dramatic improvements in ISD algorithms. To provide evidence of this claim we analyze, without loss of generality, a simplified situation. We focus on the case $n_0 = 2$, and consider the success probability of ISD algorithms when applied on LEDAcrypt schemes, searching for a row of the secret L (say, the first row), with weight $\omega = 2d_v(m_0+m_1)$.

In this case we expect to have the usual speed-up deriving from the presence of multiple low-weight codewords. However, quantifying this speed-up is not straightforward and requires cumbersome computations, since it also depends on the particular choice of the chosen set in ISD. Thus, to keep the description as general as possible and easy to follow, in this section we only focus on a single row of L. Exact computations for these quantities are performed in Sections 4 and 5. Furthermore, we only consider the probability that a chosen set T does not overlap with the support of the searched codeword. With this choice, we essentially capture the essence of all ISD algorithms. An analysis on a specific variant, with optimized parameters and requirements on the chosen set, might significantly improve the results of this section which, however, are already significant for the security of LEDAcrypt schemes.

Let $T \subseteq \{0, \cdots, n-1\}$ be a set of dimension k: for $a \sim \mathcal{U}_\omega$, we have

$$\Pr\left[T \cap \mathrm{S}(a) = \varnothing \middle| \, a \sim \mathcal{U}_\omega\right] = \frac{\binom{n-\omega}{k}}{\binom{n}{k}}.$$

Note that this probability does not depend on the particular choice of T, but just on its size. When a purely random QC-MDPC code is used, as in BIKE [22], the first row of the secret parity-check matrix is well modeled as a random sample from $\mathcal{U}_\omega$. The previous probability can also be described as the ratio between the number of n-uples of weight ω whose support is disjoint with T, and that of all possible samples from $\mathcal{U}_\omega$; in schemes such as BIKE, this also corresponds to the probability that a secret key satisfies the requirement on an arbitrary set T.

As we show in the remainder of this section, in LEDAcrypt such a fraction can actually be made significantly larger, when T is properly chosen. To each choice, we can then associate a family of *weak keys*, that is, secret keys for which the corresponding first row of L does not overlap with T. We formally define the notion of weak keys in the following.

Definition 3.1. *Let $\mathcal{K}$ be the public key space of LEDAcrypt with parameters n_0, p, d_v, m_0, m_1. Let $T \subseteq \{0, \cdots, n_0p-1\}$ of cardinality $n-k=p$ and $\mathcal{W} \subseteq \mathcal{K}$ be the set of all public keys corresponding to secret keys $sk=(H,Q)$ such that the first row in the corresponding $L=HQ$ has support that is disjoint with T. Finally, we define $\omega = n_0(m_0+m_1)d_v$ and $\mathcal{U}_\omega$ as the uniform distribution of (n_0p)-tuples with weight ω. Then, we say that $\mathcal{W}$ is a set of* weak-keys *if*

$$\Pr\left[pk \in \mathcal{W} | (sk, pk) \leftarrow \mathsf{KeyGen}()\right] \gg \Pr\left[T \cap \mathrm{S}(a) = \varnothing | a \sim \mathcal{U}_\omega\right] = \frac{\binom{n_0p-\omega}{p}}{\binom{n_0p}{p}}.$$

Roughly speaking, we have a family of weak keys when, for a specific set choice, the number of keys meeting the requirement on the support is significantly larger than the one that we would have for the uniform case. Indeed, for all such keys, we will have a strongly bias in the matrix L, since null positions can be guessed with high probability; as we describe in Sections 4 and 5, this fact opens up for strong attacks against very large portions of keys.

3.1 Preliminary considerations on sparse polynomials multiplications

We now recall some basic fact about polynomial multiplication in the rings $\mathbb{F}_2[x]/\langle x^p+1\rangle$ and $\mathbb{Z}[x]/\langle x^p-1\rangle$, which will be useful for our treatment. Let $a, b \in \mathbb{F}_2[x]/\langle x^p+1\rangle$ and $c = ab$; we then have

$$c_i = \bigoplus_{z=0}^{p-1} a_z b_{z'}, \quad z' = i - z \mod p,$$

where the operator $\bigoplus$ highlights the fact that the sum is performed over $\mathbb{F}_2$. Taking into account antisupports, we can rewrite the previous equation as

$$c_i = \bigoplus_{\substack{z \notin \bar{\mathrm{S}}(a) \\ z'=i-z \mod p,\ z' \notin \bar{\mathrm{S}}(b)}}^{p-1} a_z b_{z'}. \tag{1}$$

Let $N(a,b,i)$ denote the set of terms that contribute to the sum in Eq. (1), i.e.

$$N(a,b,i) = \left\{z \text{ s.t. } z \notin \bar{\mathrm{S}}(a) \text{ and } i - z \mod p \notin \bar{\mathrm{S}}(b)\right\}.$$

We now denote with $\tilde{a}$ and $\tilde{b}$ the polynomials obtained by lifting a and b over $\mathbb{Z}[x]/\langle x^p-1\rangle$ i.e., by mapping the coefficients of a and b into $\{0,1\} \subset \mathbb{Z}$. Let $\tilde{c} =$

$\tilde{a}\tilde{b}$: we straightforwardly have that $c \equiv \tilde{c} \mod 2$, $|N(a,b,i)| = \tilde{c}_i$ and $\sum_{i=0}^{p-1} \tilde{c}_i = \mathrm{wt}(a) \cdot \mathrm{wt}(b)$. Let $a' \in \mathbb{Z}[x]/\langle x^p + 1\rangle$ with coefficients in $\{0, 1\}$, such that $\mathrm{S}(a') \supseteq \mathrm{S}(a)$, i.e., such that its support contains that of a (or, in another words, such that its antisupport is contained in that of a); an analogous definition holds for b'. Indeed, we can write $a' = \tilde{a} + s_a$, where $s_a \in \mathbb{Z}[x]/\langle x^p + 1\rangle$ and whose i-th coefficient is equal to 0 if $a'_i = a_i$, and equal to 1 otherwise; with analogous notation, we can write $b' = \tilde{b} + s_b$. Then

$$c' = a'b' = (\tilde{a} + s_a)(\tilde{b} + s_b) = \tilde{a}\tilde{b} + s_a\tilde{b} + s_b\tilde{a} + s_a s_b = \tilde{c} + s_a\tilde{b} + s_b\tilde{a} + s_a s_b.$$

Since $s_a\tilde{b}$, $s_b\tilde{a}$ and $s_a s_b$ have all non-negative coefficients, we have

$$c'_i \geq \tilde{c}_i = |N(a,b,i)| \geq 0, \forall i \in \{0, \cdots, p-1\}. \tag{2}$$

We now derive some properties that link the coefficients of c' to those of c; as we show, knowing portions of the antisupports of a and b is enough to gather information about the coefficients in their product.

Lemma 3.2. *Let $a, b \in \mathbb{F}_2[x]/\langle x^p + 1\rangle$, and $J_a, J_b \subseteq \{0, \cdots, p-1\}$ such that $J_a \supseteq \mathrm{S}(a)$ and $J_b \supseteq \mathrm{S}(b)$. Let $a', b' \in \mathbb{Z}[x]/\langle x^p - 1\rangle$ be the polynomials whose coefficients are null, except for those indexed by J_a and J_b, respectively, which are set as 1. Let $c = ab \in \mathbb{F}_2[x]/\langle x^p + 1\rangle$ and $c' = a'b' \in \mathbb{Z}[x]/\langle x^p - 1\rangle$; then*

$$c'_i = 0 \implies c_i = 0.$$

Proof. The result immediately follows from (2) by considering that if $c'_i = 0$ then necessarily $|N(a,b,i)| = 0$ and, subsequently, $c_i = 0$. □

When the weight of $c = ab$ is maximum, i.e., equal to $\mathrm{wt}(a) \cdot \mathrm{wt}(b)$, the probability to have null coefficients in c_i can be related to the coefficients in c'_i; in analogous way, we can also derive the probability that several bits are simultaneously null. These relations are formalized in the following Lemma.

Lemma 3.3. *Let $a, b \in \mathbb{F}_2[x]/\langle x^p + 1\rangle$, with respective weights ω_a and ω_b, such that $\omega = \omega_a\omega_b \leq p$, and $c = ab$ has weight ω. Let $J_a, J_b \subseteq \{0, \cdots, p-1\}$ such that $J_a \supseteq \mathrm{S}(a)$ and $J_b \supseteq \mathrm{S}(b)$. Let $a', b' \in \mathbb{Z}[x]/\langle x^p - 1\rangle$ be the polynomials whose coefficients are null, except for those indexed by J_a and J_b, respectively, which are set as 1; finally, let $M = |J_a| \cdot |J_b|$.*

i) Let c'_i be the i-th coefficient of $c' = a'b'$; then

$$\Pr\left[c_i = 0 | c'_i\right] = \gamma(M, \omega, c'_i) = \left(1 + \omega \cdot \frac{c'_i}{M + 1 - \omega - c'_i}\right)^{-1}.$$

ii) For $V = \{v_0, \cdots, v_{t-1}\} \subseteq \{0, \cdots, p-1\}$, we have

$$\Pr\left[\mathrm{wt}(c|_V) = 0 \mid c'\right] = \zeta(V, c', \omega) = \prod_{\ell=0}^{t-1} \gamma\big(M - \textstyle\sum_{j=0}^{\ell-1} c'_{v_j}, \omega, c'_{v_\ell}\big).$$

Proof. The results follow from a combinatorial argument. See [3], Appendix B.3 for details. □

3.2 Identifying families of weak keys

We are now ready to use the results presented in the previous section to describe how, in LEDAcrypt, families of weak keys as in Def. 3.1 can be identified. We base our strategy on the results of Lemmas 3.2 and 3.3. Briefly, we guess "containers" for each polynomial in the secret key, i.e., polynomials over $\mathbb{Z}[x]/\langle x^p - 1\rangle$ whose support contains that of the corresponding polynomials in $\mathbb{F}_2[x]/\langle x^p + 1\rangle$. We then combine such containers, to find positions that, with high probability, do not point at set coefficient in the polynomials in $L = HQ$. Assuming that the initial choice for the containers is right, we can then use the results of Lemmas 3.2 and 3.3 to determine such positions. For the sake of simplicity, and without loss of generality, we describe our ideas for the practical case of $n_0 = 2$.

Operatively, to build a set T defining an eventual set of weak keys, we rely on the following procedure.

1. Consider sets J_{H_i} such that $J_{H_i} \supseteq \mathrm{S}(H_i)$, for $i = 0, 1$; the cardinality of J_{H_i} is denoted as B_{H_i}. In analogous way, define sets $J_{Q_{i,j}}$, for $i = 0, 1$ and $j = 0, 1$, with cardinalities $B_{Q_{i,j}}$.
2. To each set J_{H_i}, associate a polynomial $H'_i \in \mathbb{Z}[x]/\langle x^p - 1\rangle$, taking values in $\{0, 1\}$ and whose support corresponds to J_{H_i}; in analogous way, construct polynomials $J_{Q_{i,j}}$ from the sets $J_{Q_{i,j}}$. Compute
$$L'_{i,j} = H'_j Q'_{j,i} \in \mathbb{Z}[x]/\langle x^p - 1\rangle, \ \ (i, j) \in \{0, 1\}^2.$$
3. Compute
$$L'_i = L'_{i,0} + L'_{i,1} = H'_0 Q'_{0,i} + H'_1 Q'_{1,i} \in \mathbb{Z}[x]/\langle x^p - 1\rangle.$$
Let π_i, with $i = 0, 1$, be a permutation such that the coefficients of $\pi_i\big(L'_i\big)$ are in non decreasing order. Group the first $\left\lfloor \frac{p}{2} \right\rfloor$ entries of π_0 in a set T_0, and the first $\left\lceil \frac{p}{2} \right\rceil$ ones of π_1 in a set T_1. Define T as $T = T_0 \cup \{p + \ell |\, \ell \in T_1\}$.

A visual representation of the above constructive method to search for weak keys is described in [3], Appendix C.

Essentially, our proposed procedure to find families of weak keys starts from the sets J_{H_i} and $J_{Q_{i,j}}$, which we think of as "containers" for the secret key, i.e., sets containing the support of the corresponding polynomial in the secret key. Their products yield polynomials $L'_{i,j}$, which are containers for the products $H_i Q_{j,i}$. Because of the maximum weight requirement in LEDAcrypt key generation, each $L'_{i,j}$ matches the hypothesis required by the Lemma 3.3: the lowest entries in $L'_{i,j}$ correspond to the coefficients that, with the highest probability, are null in $H'_i Q'_{j,i}$. We remark that, because of Lemma 3.2, a null coefficient in $L'_{i,j}$ means that the corresponding coefficients in $H_j Q_{j,i}$ must be null. Finally, we need to combine the coefficients of the polynomials $L'_{i,j}$, to identify positions that are very likely to be null in each L_i. The approach we consider consists in choosing the positions that correspond to coefficients with minimum values in the sums $L'_{i,0} + L'_{i,1}$. This simple criterion is likely to be not optimal, but allows to avoid cumbersome notation and computations; furthermore, as we show next, it already detects significantly large families of weak keys.

The number of secret keys that meet the requirements on T, i.e., keys leading to polynomials L_0 and L_1 that do not overlap with the chosen sets T_0 and T_1, respectively, clearly depends on the particular choice for the containers. In the remainder of this section, we describe how such a quantity can be estimated. For the sake of simplicity, we analyze the case in which the starting sets for the containers have constant size, i.e., $B_{H_i} = B_H$ and $B_{Q_{i,j}} = B_Q$, for all i and j; furthermore, we choose $J_{H_0} = J_{H_1}$, $J_{Q_{0,0}} = J_{Q_{1,1}}$ and $J_{Q_{1,0}} = J_{Q_{0,1}}$.

First of all, let $\mathcal{J}$ be the set of secret keys whose polynomials are contained in the sets J_{H_i} and $J_{Q_{i,j}}$; the cardinality of this set can be estimated as

$$|\mathcal{J}| = \eta\left(\binom{B_H}{d_v}\binom{B_Q}{m_0}\binom{B_Q}{m_1}\right)^2,$$

where η is the acceptance ratio in key generation, i.e., the probability that a random choice of matrices H and Q leads to a matrix L with full weight.

We now estimate the number of keys in $\mathcal{J}$ that produce polynomials L_0 and L_1 corresponding to a correct choice for T_0 and T_1, i.e., such that their supports are disjoint with T_0 and T_1, respectively. For each product $H_iQ_{i,j}$, we know i) that it has full weight, not larger than p, and ii) that sets J_{H_i}, $J_{Q_{i,j}}$ are containers for H_i and $Q_{i,j}$, respectively. Then, Lemma 3.3 can be used to estimate the portion of valid keys. For instance, we consider the polynomial $L_0 = H_0Q_{0,0} + H_1Q_{1,0}$: the coefficients that are indexed by T_0 will be null when both the supports of $H_0Q_{0,0}$ and $H_1Q_{1,0}$ are disjoint with T_0. If we neglect the fact that these two products are actually correlated (because of the full weight requirement on L_0), then the probability that L_0 does not overlap with T_0, which we denote as $\Pr\left[\texttt{null}(T_0)\right]$, is obtained as

$$\Pr\left[\texttt{null}(T_0)\right] = \zeta\big(T_0, L'_{0,0}, m_0d_v\big) \cdot \zeta\big(T_0, L'_{0,1}, m_1d_v\big),$$

where ζ is defined in Lemma 3.3. The above quantity can then be used to estimate the fraction of keys in $\mathcal{J}$ for which the support of L_0 does not overlap with T_0; we remark that, as highlighted by the above formula, this quantity strongly depends on the choices on J_{H_0}, J_{H_1}, $J_{Q_{0,0}}$, $J_{Q_{1,0}}$.
With the same reasoning, and with analogous notation, we compute $\Pr\left[\texttt{null}(T_1)\right]$; because of the simplifying restrictions on $J_{Q_{i,j}}$, this probability is equal to $\Pr\left[\texttt{null}(T_0)\right]$.

Then, if we neglect the correlation between L_0 and L_1 (since H_0 and H_1 are involved in the computation of both polynomials), the probability that a random key from $\mathcal{J}$ is associated to a valid L, i.e., that it leads to polynomials L_0 and

L_1 that respectively do not overlap with T_0 and T_1, can be estimated as

$$\begin{aligned}\Pr\left[\texttt{null}(T)\right] &= \Pr\left[\texttt{null}(T_0)\right] \cdot \Pr\left[\texttt{null}(T_1)\right] \\ &= \big(\Pr\left[\texttt{null}(T_0)\right]\big)^2 \\ &= \Big(\zeta\big(T_0, L'_{0,0}, m_0 d_v\big) \cdot \zeta\big(T_0, L'_{0,1}, m_1 d_v\big)\Big)^2.\end{aligned}$$

Thus we conclude that the number of keys whose polynomials are contained by the chosen sets, and such that the corresponding L does not overlap with T, can be estimated as $|\mathcal{J}| \cdot \Pr[\texttt{null}(T)]$.
Then, for the set of secret keys where T does not intercept the first row of L, which we denote with $\mathcal{W}$, we have

$$|\mathcal{W}| \geq |\mathcal{J}| \cdot \Pr[\texttt{null}(T)]. \tag{3}$$

The inequality comes from the fact the right term in the above formula only counts keys with polynomials contained by the initially chosen sets; even if such property is not satisfied, it may still happen that the resulting L does not overlap with T (thus, we are underestimating the cardinality of $\mathcal{W}$).

3.3 Results

In this section we provide practical examples on choices for containing sets, leading to actual families of weak keys. To this end, we need to define clear criteria on how the sets J_{H_i} and $J_{Q_{i,j}}$ can be selected. For the sake of simplicity, we restrict our attention to the cases $J_{H_0} = J_{H_1} = J_H$ and $J_{Q_{0,0}} = J_{Q_{0,1}} = J_{Q_{1,0}} = J_{Q_{1,1}} = J_Q$. We here consider two different strategies to pick these sets.

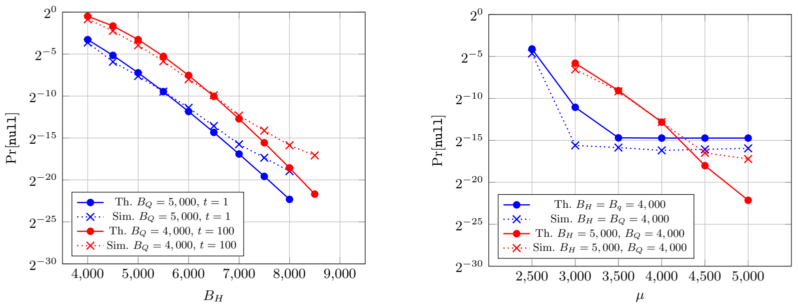
- *Type I*: for $i = 0, 1$, $\delta \in \{0, \cdots, p-1\}$ and $t \in \{1, \cdots, p-1\}$, we choose

$$\begin{aligned} J_H &= \{\ell t \mod p \,|0 \leq \ell \leq B_H - 1\}, \\ J_Q &= \{\delta + \ell t \mod p \,|0 \leq \ell \leq B_Q - 1\}.\end{aligned}$$

- *Type II*: for $i = 0, 1$, we choose $J_{H_0} = J_{H_1}$ as the union of disjoint sets, formed by contiguous positions. Analogous choice is adopted for J_Q.

To provide numerical evidences for our analysis, in Figure 1 we compare the simulated values of $\Pr[\texttt{null}(T)]$ with the ones obtained with theoretical expression, for parameters of practical interest and for some Types I and II choices. The simulated probabilities have been obtained by generating random secret keys from $\mathcal{J}$ and, as our results show, are well approximated by the theoretical expression. This shows that Eq. 3 provides a good estimate for the fraction of keys in $\mathcal{J}$ that meet the requirement on the corresponding set T.

Tables 2, 3 display results testing various weak key families of Type I and II, for two different LEDAcrypt parameters sets. According to the reasoning in the previous section, the values reported in the last column can be considered as a rough (and likely conservative) estimate for the probability that a random key belongs to the corresponding set $\mathcal{W}$. Our results show that the identified families of keys meet Definition **??**, so can actually be considered weak.



(a) Type I

(b) Type II

Fig. 1. Comparison between simulated and theoretical values for Pr[`null`], for $p = 14939$, $d_v = 11$, $m_0 = 4$, $m_1 = 3$. The values reported in Figure (a) are all referred to the case $\delta = 0$. In Figure (b), the blue curves correspond to the choice $J_H = J_Q = \{0, \cdots, 1999\} \cup \{\mu, \cdots, \mu + 1999\}$, while the red curves correspond to $J_H = \{0, \cdots, 2499\} \cup \{\mu, \cdots, \mu + 2499\}$ and $J_Q = \{0, \cdots, 3999\}$.

Type	Family Parameters	$\frac{\lvert\mathcal{J}\rvert \cdot \Pr[\texttt{null}(T)]}{\lvert\mathcal{K}\rvert}$
I	$B_H = B_Q = 7470$ $\delta = 0, t = 1$	$2^{-99.88}$
I	$B_H = 8000, B_Q = 4000$ $\delta = 2000, t = 1$	$2^{-85.25}$
I	$B_H = 8500, B_Q = 4000$ $\delta = 0, t = 127$	$2^{-90.23}$
II	$J_H = \{0, \cdots, 4499\} \cup \{7000, \cdots, 11499\}$ $J_Q = \{0, \cdots, 2499\} \cup \{8000, \cdots, 10499\}$	$2^{-101.53}$

Table 2. Fraction of weak keys, for LEDAcrypt instances designed for 128-bit security, with parameters $n_0 = 2$, $p = 14939$, $d_v = 11$, $m_0 = 4$, $m_1 = 3$, for which $\eta \approx 0.7090$. For this parameter set, probability of randomly guessing a null set of dimension p, in a vector of length $2p$ and weight $2(m_0 + m_1)d_v$, is $2^{-154.57}$.

Remark 1. The results we have shown in this section only represent a qualitative evidence of the existence of families of weak keys in LEDAcrypt. There may exist many more families of weak keys, having a complete different structure from the ones we have studied. Additionally, the parameters we have considered for types I and II may not be the optimal ones, but already identify families of weak keys. In the next sections we provide a detailed analysis for families of keys of type I and II, and furthermore specify the actual complexity of a full cryptanalysis exploiting such a key structure.

Type	Family Parameters	$\frac{\lvert\mathcal{J}\rvert\cdot\Pr[\texttt{null}(T)]}{\lvert\mathcal{K}\rvert}$
I	$B_H = 18000, B_Q = 9000$ $\delta = 9000, t = 1$	$2^{-125.18}$
I	$B_H = 24000, B_Q = 12000$ $\delta = 0, t = 1$	$2^{-184.21}$
I	$B_H = 18000, B_Q = 9000$ $\delta = 0, t = 5$	$2^{-125.18}$
II	$J_H = \{0, \cdots, 20999\}$ $J_Q = \{0, \cdots, 3999\} \cup \{10000, \cdots, 13999\} \cup \{20000, \cdots, 23999\}$	$2^{-270.30}$

Table 3. Fraction of weak keys, for LEDAcrypt instances designed for 256-bit security, with parameters $n_0 = 2$, $p = 36877$, $d_v = 11$, $m_0 = 7$, $m_1 = 6$, for which $\eta \approx 0.614$. For this parameter set, probability of randomly guessing a null set of dimension p, in a vector of length $2p$ and weight $2(m_0 + m_1)d_v$, is $2^{-286.80}$.

4 Explicit Attack on the Weakest Class of Keys

In the previous section we described how the product structure in LEDAcrypt leads to an highly biased distribution in set positions in L. As we have hinted, this property may be exploited to improve cryptanalysis techniques based on ISD algorithms. In this section, we present an attack against a class of weak keys in LEDAcrypt's design. We begin by identifying what appear to be the weakest class of keys (though large enough in number to constitute a serious, practical problem for LEDAcrypt). It is easily seen that the class of keys we consider in this section corresponds to a particular case of type I, introduced in Section 3.3. We proceed to provide a simple, single-iteration ISD algorithm to recover these keys, then analyze the fraction of all of LEDAcrypt's keys that would be recovered by this attack. Afterward, we show how to extend the ISD algorithm to more than one iteration, so as to enlarge the set of keys recovered by a similar enough of effort per key. We conclude by considering the effect of advanced ISD algorithms on the attack as well as the relationship between the rejection sampling step in LEDAcrypt's KeyGen and our restriction to attacking a subspace of the total key space.

4.1 Attacking an example (sub)class of ultra-weak keys

The simplest and, where it works, most powerful version of the attack dramatically speeds up ISD for a class of ultra-weak keys chosen under parameter sets where $n_0 = 2$. One example (sub)class of ultra-weak keys are those keys where the polynomials L_0 and L_1 are of degree at most $\frac{p}{2}$. Such keys can be found by a single iteration of a very simple ISD algorithm. We describe this simple attack as follows.

The attacker chooses the information set to consist of the last $\frac{p-1}{2}$ columns of the first block of M and the last $\frac{p+1}{2}$ columns of the second block. If the key being attacked is one of these weak keys, the attacker can correctly guess the top row of L as being identically zero within the information set and linearly solve for the nonzero linear combination of the rows of M meeting this condition. The cost of the attack is one iteration of an ISD algorithm.

A sufficient condition for this class of weak key to occur is for the polynomials H_0, H_1, $Q_{0,0}$, $Q_{0,1}$, $Q_{1,0}$, and $Q_{1,1}$ to have degree no more than $\frac{p}{4}$. Since each of the $2m_0+2m_1+2d_v$ nonzero coefficients of these polynomials has a $\frac{1}{4}$ probability of being chosen with degree less than $\frac{p}{4}$, these weak keys represent at least 1 part in $4^{2m_0+2m_1+2d_v}$ of the key space.

4.2 Enumerating ultra-weak keys for a single information set

In fact, there are significantly more weak keys than this that can be recovered by the basic, one-iteration ISD algorithm using the information set described above. Intuitively, this is for two reasons:

1. **Equivalent keys:** There are p^2 private keys, not of this same, basic form, which nonetheless produce the same public key.
2. **Different degree constraints:** The support of the top row of L will also fall entirely outside the information set if the degree of H_0 is less than $\frac{p}{4} - \delta$ and the degrees of $Q_{0,0}$ and $Q_{0,1}$ are both less than $\frac{p}{4} + \delta$ for any $\delta \in \mathbb{Z}$ such that $-\frac{p}{4} < \delta < \frac{p}{4}$. Likewise for H_1 and $Q_{1,0}$ and $Q_{1,1}$, for a total of p keys.

Concretely, we derive the number of distinct private keys that are recovered by the one-iteration ISD algorithm in the following theorem.

Remark 2. There are p columns of each block of M. For the sake of simplicity, instead of referring to pairs of $\frac{p-1}{2}$ and $\frac{p+1}{2}$ columns, we instead use $\frac{p}{2}$ for both cases. This has a negligible effect on our results.

Theorem 4.1. *The number of distinct private keys that can be found in a single iteration of the decoding algorithm described above (where the information set is chosen to consist of the last $\frac{p}{2}$ columns of each block of M) is*

$$\begin{aligned}
&p^3 \cdot \sum_{A_0=d_v-1}^{\frac{p}{2}} \sum_{A_1=d_v-1}^{\frac{p}{2}} \left(\binom{A_0-1}{d_v-2} \binom{A_1-1}{d_v-2} \right. \\
&\cdot \left(\binom{\frac{p}{2}-A_0-2}{m_0-1} \binom{\frac{p}{2}-A_0-1}{m_1} \binom{\frac{p}{2}-A_1-1}{m_1} \binom{\frac{p}{2}-A_1-1}{m_0} \right. \\
&+ \binom{\frac{p}{2}-A_0-1}{m_0} \binom{\frac{p}{2}-A_0-2}{m_1-1} \binom{\frac{p}{2}-A_1-1}{m_1} \binom{\frac{p}{2}-A_1-1}{m_0} \\
&+ \binom{\frac{p}{2}-A_0-1}{m_0} \binom{\frac{p}{2}-A_0-1}{m_1} \binom{\frac{p}{2}-A_1-2}{m_1-1} \binom{\frac{p}{2}-A_1-1}{m_0} \\
&\left. \left. + \binom{\frac{p}{2}-A_0-1}{m_0} \binom{\frac{p}{2}-A_0-1}{m_1} \binom{\frac{p}{2}-A_1-1}{m_1} \binom{\frac{p}{2}-A_1-2}{m_0-1} \right) \right) \\
&\cdot \left(1 - O\left(\frac{m}{p} \right) \right).
\end{aligned} \tag{4}$$

Proof. We count the number of ultra-weak keys as follows. By assumption, all nonzero bits in each block of an ultra-weak key are contained in some consecutive stretch of size $\leq \frac{p}{2}$. Thus these ultra-weak keys contain a stretch of at least $\frac{p}{2}$ zero bits. This property applies directly to the polynomials $H_0Q_{0,0} + H_1Q_{1,0}$ and $H_0Q_{0,1} + H_1Q_{1,1}$, and must also hold for H_0 and H_1. We index the number of ultra-weak keys according to the first nonzero coefficient of these polynomials after the stretch of zero bits in cyclic ordering.

We begin by considering H, Q though not requiring HQ to have full weight. We are using an information set consisting of the same columns for both $H_0Q_{0,0}+ H_1Q_{1,0}$ and $H_0Q_{0,1} + H_1Q_{1,1}$. Therefore we count according the first nonzero bit of the sum $H_0Q_{0,0} + H_1Q_{1,0} + H_0Q_{0,1} + H_1Q_{1,1}$. Let l be the location of the first nonzero bit of this sum.

Let j_0, j_1 be the locations of the first nonzero bit of H_0, H_1, respectively. Suppose that the nonzero bits of H_0, H_1 are located within a block of length A_0, A_1, respectively.

By LEDAcrypt's design, $d_v \leq A_i, i \in \{0, 1\}$ and by assumption on the chosen information set, $A_i \leq \frac{p}{2}, i \in \{0, 1\}$. Once j_0 is fixed, there are $\sum_{A_0=d_v-1}^{\frac{p}{2}} \binom{A_0-1}{d_v-2}$ ways to arrange the remaining bits of H_0. Thus there are

$$\sum_{j_0=1}^{p-1} \sum_{A_0=d_v-1}^{\frac{p}{2}} \binom{A_0-1}{d_v-2} \sum_{j_1=1}^{p-1} \sum_{A_1=d_v-1}^{\frac{p}{2}} \binom{A_1-1}{d_v-2} \tag{5}$$

many bit arrangements of H_0, H_1.

Once j_0, j_1 are fixed, there are four blocks of Q which may influence the location l. We compute the probability that only one block of Q may influence l at a time.

If l is influenced by $Q_{0,0}$, there are $\binom{\frac{p}{2}-A_0-2}{m_0-1}$ ways the remaining bits of $Q_{0,0}$ can fall, $\binom{\frac{p}{2}-A_0-1}{m_1}$ arrangements of the bits of $Q_{0,1}$, $\binom{\frac{p}{2}-A_1-1}{m_1}$ arrangements of the bits of $Q_{1,0}$, and $\binom{\frac{p}{2}-A_1-1}{m_0}$ arrangements of the bits of $Q_{1,1}$. If l is influenced by $Q_{0,1}$, there are $\binom{\frac{p}{2}-A_0-2}{m_0}$ arrangements of the bits of $Q_{0,0}$, $\binom{\frac{p}{2}-A_0-1}{m_1-1}$ ways the remaining bits of $Q_{0,1}$ can fall, $\binom{\frac{p}{2}-A_1-1}{m_1}$ arrangements of the bits of $Q_{1,0}$, and $\binom{\frac{p}{2}-A_1-1}{m_0}$ arrangements of the bits of $Q_{1,1}$. Similar estimates hold for $Q_{1,0}$, or $Q_{1,1}$.

We sum over the l locations considering each of the blocks of Q and their respective weights. Then the overall sum is

$$\begin{aligned}
&\sum_{j_0=0}^{p-1}\sum_{A_0=d_v-1}^{\frac{p}{2}}\binom{A_0-1}{d_v-2}\sum_{j_1=0}^{p-1}\sum_{A_1=d_v-1}^{\frac{p}{2}}\binom{A_1-1}{d_v-2}\\
&\cdot\sum_{l=0}^{p-1}\left(\binom{\frac{p}{2}-A_0-2}{m_0-1}\binom{\frac{p}{2}-A_0-1}{m_1}\binom{\frac{p}{2}-A_1-1}{m_1}\binom{\frac{p}{2}-A_1-1}{m_0}\right.\\
&+\binom{\frac{p}{2}-A_0-1}{m_0}\binom{\frac{p}{2}-A_0-2}{m_1-1}\binom{\frac{p}{2}-A_1-1}{m_1}\binom{\frac{p}{2}-A_1-1}{m_0}\\
&+\binom{\frac{p}{2}-A_0-1}{m_0}\binom{\frac{p}{2}-A_0-1}{m_1}\binom{\frac{p}{2}-A_1-2}{m_1-1}\binom{\frac{p}{2}-A_1-1}{m_0}\\
&\left.+\binom{\frac{p}{2}-A_0-1}{m_0}\binom{\frac{p}{2}-A_0-1}{m_1}\binom{\frac{p}{2}-A_1-1}{m_1}\binom{\frac{p}{2}-A_1-2}{m_0-1}\right)\\
&\cdot\left(1-O\left(\frac{m}{p}\right)\right).
\end{aligned} \tag{6}$$

Failure to impose full weight requirements on HQ introduces double-counting. This occurs when more than one block of Q influences l, though the probability of this event will not exceed $O(\frac{m}{p})$. The constant sums yield the factor of p^3. □

We can now estimate the percentage of these recovered, ultra-weak keys out of all possible keys.

Theorem 4.2. *Let $m = m_0 + m_1, x = \frac{A_0}{p}, y = \frac{A_1}{p}$. Out of $\binom{p}{d_v}^2\binom{p}{m_0}^2\binom{p}{m_1}^2$ possible keys, we estimate the percentage of ultra-weak keys found in a single iteration of the decoding algorithm above as*

$$d_v{}^2(d_v-1)^2 m\int_{x=0}^{\frac{1}{2}}\int_{y=0}^{\frac{1}{2}}(xy)^{d_v-2}\left(\left(\frac{1}{2}-x\right)\left(\frac{1}{2}-y\right)\right)^m\left(\frac{1}{\frac{1}{2}-x}+\frac{1}{\frac{1}{2}-y}\right)\mathrm{d}y\mathrm{d}x.$$

Proof. Note that the lines $2-5$ of (4) are approximately

$$\binom{\frac{p}{2}-A_0}{m_0}\binom{\frac{p}{2}-A_0}{m_1}\binom{\frac{p}{2}-A_1}{m_1}\binom{\frac{p}{2}-A_1}{m_0}\left(\frac{m_0+m_1}{\frac{p}{2}-A_1}+\frac{m_0+m_1}{\frac{p}{2}-A_0}\right). \tag{7}$$

For $b, c \in \{0, 1\}$,

$$\binom{\frac{p}{2}-A_b}{m_c}\approx\binom{p}{m_c}\left(\frac{1}{2}-\frac{A_b}{p}\right)^{m_c} \tag{8}$$

and

$$\binom{A_b-1}{d_v-2}\approx\binom{p}{d_v-2}\left(\frac{A_b}{p}\right)^{d_v-2} \tag{9}$$

since p is much larger than m_0, m_1, d_v. We rewrite (4) using the approximations of expressions (7,8) as

$$p^3 \sum_{A_0=d_v-1}^{\frac{p}{2}} \binom{A_0-1}{d_v-2} \sum_{A_1=d_v-1}^{\frac{p}{2}} \binom{A_1-1}{d_v-2} \binom{p}{m_0}^2 \left(\frac{1}{2}-\frac{A_0}{p}\right)^{m_0+m_1} \tag{10}$$

$$\binom{p}{m_1}^2 \left(\frac{1}{2}-\frac{A_1}{p}\right)^{m_0+m_1} \left(\frac{m_0+m_1}{\frac{p}{2}-A_1}+\frac{m_0+m_1}{\frac{p}{2}-A_0}\right). \tag{11}$$

Applying approximation (9) further reduces expression (10) to

$$\begin{aligned}
&p^3 \binom{p}{m_0}^2 \binom{p}{m_1}^2 \binom{p}{d_v-2}^2 \sum_{A_0=d_v-1}^{\frac{p}{2}} \left(\frac{A_0}{p}\right)^{d_v-2} \sum_{A_1=d_v-1}^{\frac{p}{2}} \left(\frac{A_1}{p}\right)^{d_v-2} \\
&\quad \left(\frac{1}{2}-\frac{A_0}{p}\right)^{m_0+m_1} \left(\frac{1}{2}-\frac{A_1}{p}\right)^{m_0+m_1} \left(\frac{m_0+m_1}{\frac{p}{2}-A_1}+\frac{m_0+m_1}{\frac{p}{2}-A_0}\right) \\
=&p^2 \binom{p}{d_v-2}^2 \binom{p}{m_0}^2 \binom{p}{m_1}^2 m \sum_{A_0=d_v-1}^{\frac{p}{2}} \sum_{A_1=d_v-1}^{\frac{p}{2}} \left(\frac{A_0}{p}\frac{A_1}{p}\right)^{d_v-2} \left(\frac{1}{2}-\frac{A_0}{p}\right)^{m} \\
&\quad \left(\frac{1}{2}-\frac{A_1}{p}\right)^{m} \left(\frac{1}{\frac{1}{2}-\frac{A_0}{p}}+\frac{1}{\frac{1}{2}-\frac{A_1}{p}}\right).
\end{aligned}$$

Letting $x=\frac{A_0}{p}, y=\frac{A_1}{p}$, this is approximated by

$$\begin{aligned}
&p^2 \binom{p}{d_v}^2 \binom{p}{m_0}^2 \binom{p}{m_1}^2 m \frac{{d_v}^2(d_v-1)^2}{(p-d_v+2)^2(p-d_v+1)^2} \\
&\cdot p^2 \int_{x=0}^{\frac{1}{2}} \int_{y=0}^{\frac{1}{2}} (xy)^{d_v-2} \left(\frac{1}{2}-x\right)^m \left(\frac{1}{2}-y\right)^m \left(\frac{1}{\frac{1}{2}-x}+\frac{1}{\frac{1}{2}-y}\right) \mathrm{d}y\mathrm{d}x.
\end{aligned}$$

Dividing by $\binom{p}{d_v}^2\binom{p}{m_0}^2\binom{p}{m_1}^2$, the result follows. □

Evaluating this percentage with the claimed-256-bit ephemeral (CPA-secure) key parameters of LEDAcrypt — $d_v = 11, m = 13$ — we determine that 1 in $2^{72.8}$ ephemeral keys are broken by one iteration of ISD. Similarly for the long-term (CCA-secure) key setting, we evaluate with the claimed 256-bit parameters — $d_v = 13, m = 13$ — and conclude the number of long-term keys broken is 1 in $2^{80.6}$.

This result merely determines the number of keys that can be recovered given that the information set of both blocks of M is chosen to be the last $\frac{p}{2}$ columns.[2] In the following, we turn to demonstrating a class of additional information sets that are as effective as this one.

[2] For the reader, we point out that if, hypothetically, we had a sufficiently large number of *totally independent* information sets that were equally "rewarding" in recovering keys, this would straightforwardly imply $\approx 2^{72.8}$-time and $\approx 2^{80.6}$-time "full" attacks against LEDAcrypt's claimed-256-bit parameters rather than weak-key attacks.

Remark 3. We remind the reader that instead of referring to the pairs of $\frac{p-1}{2}$, $\frac{p+1}{2}$ columns of blocks of M, we use $\frac{p}{2}$ in both cases. This has a negligible effect on our results.

4.3 Enumerating ultra-weak keys for all information sets

Now we will demonstrate a multi-iteration ISD attack that is effective against the class of all ultra-weak keys. To set up the discussion, we begin by highlighting two, further "degrees of freedom," which will allow us to find additional, relevant information sets to guess:

1. **Changing the ring representation:** Contiguity of indices depends on the choice of ring representation. The large family of ring isomorphisms on $\mathbb{Z}[x]/\langle x^p - 1\rangle$ given by $f(x) \to f(x^t)$ for $t \in [0, p]$ preserves Hamming weight. For example, we can use the family of polynomials

   $$H'_i = Q'_{i,j} = 1 + x^t + x^{2t} + ... + x^{\lfloor \frac{p}{4} \rfloor t}$$

   in this attack, since there exists one t such that H'_i has consecutive nonzero coefficients. Choices of $t \in \{1, \ldots, \frac{p-1}{2}\}$ yield independent information sets (noting that choices of t and $-t \mod p$ yield equivalent information sets).

2. **Changing the relative offset of the two consecutive blocks:** We can also change the beginning index of the consecutive blocks produced within L'_0 or L'_1 (by modifying the beginning indices of H'_i and $Q'_{i,j}$ to suit). Note that shifting both L'_0 and L'_1 by the same offset will recover equivalent keys. However, if we fix the beginning index of L'_0 and allow the beginning index of L'_1 to vary, we can find more, mostly independent information sets in order to recover more, distinct keys. The exact calculation of how far one should shift L'_1's indices for a practically effective attack is somewhat complex; we perform this analysis below in the remainder of this subsection.

Recall that in the prior 1-iteration attack, we considered *one* example class of ultra-weak keys – namely, those keys where the polynomials L_0 and L_1 are of degree at most $\frac{p}{2}$. Here, we will now take a broader view on the weakest-possible keys.

Definition 4.3. *We define the* **class of ultra-weak keys** *to be those where, in some ring representation, both* $H_0Q_{0,0} + H_1Q_{1,0}$ *and* $H_0Q_{0,1} + H_1Q_{1,1}$ *have nonzero coefficients that lie within a block of* $\frac{p-1}{2}$*-many consecutive (modulo p) degrees.*

Our goal will be now to find a multi-iteration ISD algorithm — by estimating how far to shift the offset of L'_1 per iteration — that recovers as much of the class of ultra-weak keys as possible without "overly wasting" the attacker's computational budget. Toward this end, recall that we have a good estimate in Theorem 4.2 of the fraction of keys ($2^{-72.8}$, resp. $2^{-80.6}$) recovered by the best-case, single iteration of our ISD algorithm. In what follows, we will first calculate the fraction of ultra-weak keys as a part of the total key space.

Let 2^{-X} be the fraction of all keys recovered by the best-case, single iteration of our previous ISD algorithm. Let 2^{-Y} be the fraction of ultra-weak keys among all keys. On the assumption that every ring representation leads to independent information sets (chosen uniformly for each invocation of ISD) and on the assumption that independence of ISD key-recovery is maximized by shifting "as far as possible," we will compute an estimate of the number of index-shifts that should be performed by the optimal ultra-weak-key attacker as $2^Z = 2^{X-Y}$. Beyond 2^Z shifts per guess (but not until), the attacker should begin to experience diminishing returns in how many keys are recovered per shifted guess.

Therefore, given an index beginning at 1 out of p positions, the attacker will shift by $\frac{p(\frac{p-1}{2})}{2^Z}$ indices at each invocation (where the factor $\frac{p-1}{2}$ accounts for the effect of the different possible ring representations). By assumption, each such guess will be sufficiently independent to recover as many keys in expectation as the initial, best-guess case described by the 1-iteration algorithm. We note that additional, ultra-weak keys will certainly be obtained by performing more work — specifically by shifting less than $\frac{p(\frac{p-1}{2})}{2^Z}$ per guess — but necessarily at a reduced rate of reward per guess.

Toward this end, we now calculate the number of ultra-weak keys then the fraction of ultra-weak keys among all keys following the format of the previous calculation.

Theorem 4.4. *The total number of ultra-weak keys is*

$$\frac{p-1}{2}p^2 \sum_{A_0=d_v-1}^{\frac{p}{2}} \sum_{A_1=d_v-1}^{\frac{p}{2}} \binom{A_0-1}{d_v-2}\binom{A_1-1}{d_v-2} \tag{12}$$

$$\cdot \sum_{l_0=0}^{p-1} \left(\binom{\frac{p}{2}-A_0-1}{m_0-1}\binom{\frac{p}{2}-A_1-1}{m_1} + \binom{\frac{p}{2}-A_0-1}{m_0}\binom{\frac{p}{2}-A_1-1}{m_1-1} \right) \tag{13}$$

$$\cdot \sum_{l_1=0}^{p-1} \left(\binom{\frac{p}{2}-A_0-1}{m_0}\binom{\frac{p}{2}-A_1-1}{m_1-1} + \binom{\frac{p}{2}-A_0-1}{m_0-1}\binom{\frac{p}{2}-A_1-1}{m_0} \right). \tag{14}$$

Proof. The proof technique follows as in Theorem 4.1. Details are found in [3], B.1. □

Theorem 4.5. *Let $m = m_0 + m_1, x = \frac{A_0}{p}, y = \frac{A_1}{p}$. The fraction of ultra-weak keys out of all possible keys is*

$$\frac{p-1}{2}{d_v}^2(d_v-1)^2 \int_{x=0}^{\frac{1}{2}} \int_{y=0}^{\frac{1}{2}} x^{d_v-2}y^{d_v-2}\left(\frac{1}{2}-x\right)^m\left(\frac{1}{2}-y\right)^m$$
$$\left(\frac{{m_0}^2+{m_1}^2}{(\frac{1}{2}-x)(\frac{1}{2}-y)} + \frac{m_0m_1}{(\frac{1}{2}-x)^2} + \frac{m_0m_1}{(\frac{1}{2}-y)^2}\right) \mathrm{d}y\mathrm{d}x.$$

Proof. Similar techniques apply. See [3], B.2 for details. □

We evaluate the fraction of weak keys using the claimed CPA-secure parameters $p = 36877, m = 13, d_v = 11$ and determine that 1 in $2^{54.1}$ ephemeral keys are broken. Evaluating with one of the CCA-secure parameter sets $p = 152267, m = 13, d_v = 13$, approximately 1 in $2^{59.7}$ long-term keys are broken.

Given the above, we can make an estimate as to the optimal shift-distance per ISD invocation as $\frac{36{,}877(\frac{36{,}876}{2})}{2^{72.8-54.1}} \approx 1597 \approx 2^{10.6}$ for the ephemeral key parameters and $\frac{152{,}267(\frac{152{,}266}{2})}{2^{80.6-59.7}} \approx 5925 \approx 2^{12.5}$ for the long-term key parameters.

The multi-iteration ISD algorithm against the class of ultra-weak keys, then, makes its first guess (except, one in each ring representation) as in the case of the 1-iteration ISD algorithm. It then shifts the relative offset of the two consecutive blocks by the values calculated above and repeats (again, in each ring representation).

This will not recover all ultra-weak keys, but it will recover a significant fraction of them. In particular, if the support of each block of L, rather than fitting in $\frac{p}{2}$ consecutive bits fits in blocks that are smaller by at least $\frac{1}{4}$ of the shift distance. We can therefore lower bound the fraction of recovered keys by replacing factors of $\frac{1}{2}$ with factors of $\frac{p}{2}$ minus half or a quarter of the offset, all divided by p, to find the sizes of sets of private keys of which we are guaranteed to recover all, or at least half of respectively.

The multi-iteration ISD algorithm attacking the ephemeral key parameters will make $2^{72.8-54.1} \approx 2^{18.7}$ independent guesses and recover at least 1 in $2^{56.0}$ of the total keys. The multi-iteration ISD algorithm attacking the long-term key parameters will make $2^{80.6-59.7} \approx 2^{20.9}$ independent guesses and recover at least 1 in $2^{61.6}$ of the total keys.

4.4 Estimating the effect of more advanced information-set decoding

Our attempts to enumerate all weak keys were based on the assumption that the adversary was using an ISD variant that required a row of L to be uniformly 0 on all columns of the information set. The state of the art in information set decoding still allows the adversary to decode provided that a row of L has weight no more than about 6 on the information set. For example, Stern's algorithm [28] with parameter 3 would attempt to find a low weight row of L as follows.

The information set is divided into two disjoint sets of $\frac{p}{2}$ columns. The first row of L to be recovered should have weight at most 3 within each of the two sets. Further, the same row of L should have have $\Omega(\log(p))$ many consecutive 0's in column-indices that are disjoint from those of the information set. If both of these conditions occur, then a matrix inversion is performed (even though 6 non-zero bits were contained in the information set).

Note that for reasonably large p, nearly a third of the sparse vectors having weight 6 in the information set will meet both conditions. The most expensive steps in the Stern's algorithm iteration are a matrix inversion of size p and a claw finding on functions with logarithmic cost in p and domain sizes of $\binom{\frac{p}{2}}{3}$. The claw finding step is similar in cost to the matrix inversion, both having computational cost $\approx p^3$. The matrix inversion step is present in all ISD algorithms. Therefore with Stern's algorithm we can recover in a single iteration with similar cost to a single iteration of a simpler ISD algorithm, $O(1)$ of the private keys where a row of L has weight no more than 6 on the information set columns.

Recall that we choose the information set to be of size $\approx \frac{p}{2}$ in L'. The distribution of the non-zero coordinates within a successful guess of information set will be more heavily weighted toward the middle of the set and approximately triangular shaped (since these coordinates are produced by convolutions of polynomials). In particular, we will *heuristically* model both of the tails of the distribution as small triangles containing 3 bits on the left side and three bits on the right that are missed by the choice of information set.

Let $W = 2d_v(m_0 + m_1)$ denote the number of non-zero bits in L'. Then the actual fraction ϵ that the information set (in the context of advanced information set decoding) should target within L, rather than $1/2$, can be estimated by geometric area as

$$\epsilon \cdot \left(1 - \sqrt{\frac{3}{W/2}}\right) = \frac{1}{2}$$

or, re-writing:

$$\epsilon = \frac{1}{2\left(1 - \sqrt{\frac{3}{W/2}}\right)}.$$

For the claimed-256-bit ephemeral key parameters, we have $W_{\mathsf{CPA}} = 286$. For the claimed-256-bit long-term key parameters, we have $W_{\mathsf{CCA}} = 338$. Therefore,

$$\epsilon_{\mathsf{CPA}} = \frac{1}{2\left(1 - \sqrt{\frac{3}{286/2}}\right)} \approx 0.585.$$

$$\epsilon_{\mathsf{CCA}} = \frac{1}{2\left(1 - \sqrt{\frac{3}{338/2}}\right)} \approx 0.577.$$

So – heuristically – we can model the effect of using advanced information set decoding algorithms by replacing the $\frac{1}{2}$'s in the calculations of the theorems earlier in this section by ϵ_{CPA} or ϵ_{CCA} respectively.

4.5 Rejection sampling considerations

We recall that LEDACrypt's KeyGen algorithm explicitly requires that the parity check matrix L be *full weight*. Intuitively full weight means that no cancellations occur in the additions or the multiplications that are used to generate L from H and Q. Formally, the full weight condition on L can be stated as:

$$\forall i \in \{0, \ldots, n_0 - 1\},\ \ \mathsf{weight}(L_i) = d_v \sum_{j=0}^{n_0-1} m_j.$$

When a weak key notion causes rejections to occur significantly more often for weak keys than non-weak keys, we will effectively reduce the probability of weak key generation compared to our previous analysis. As an extreme example, if, for a given weak key notion, rejection sampling rejects all weak keys, then no weak keys will ever be sampled. We therefore seek to measure the probability of key rejection for both weak keys and keys in general in order to determine whether the effectiveness of this attack is reduced via rejection sampling.

Let $\mathcal{K}$, $\mathcal{W} \subset \mathcal{K}$, and KeyGen be the public key space, the weak key space, and the key generation algorithm of LEDACrypt, respectively. Let $\mathcal{K}'$, $\mathcal{W}' \subset \mathcal{K}'$, and $\mathsf{KeyGen'}$ be the associated objects if rejection sampling were omitted from LEDACrypt. We observe that since KeyGen samples uniformly from $\mathcal{K}$,

$$\Pr\left[pk \in \mathcal{W} | (pk, sk) \leftarrow \mathsf{KeyGen}()\right] = \frac{|\mathcal{W}|}{|\mathcal{K}|}.$$

This equality additionally holds when rejection sampling does not occur. Since, until now, all of our analysis has ignored rejection sampling we have effectively been measuring $|\mathcal{W}'|/|\mathcal{K}'|$. We therefore seek to find a relation that allows us determine $|\mathcal{W}|/|\mathcal{K}|$ from $|\mathcal{K}'|$ and $\mathcal{W}'|$. We observe that

$$\frac{|\mathcal{W}|}{|\mathcal{K}|} = \frac{|\mathcal{W}|}{|\mathcal{K}|}\frac{|\mathcal{W}'|}{|\mathcal{W}'|}\frac{|\mathcal{K}'|}{|\mathcal{K}'|} = \frac{|\mathcal{W}'|}{|\mathcal{K}'|}\frac{|\mathcal{W}|}{|\mathcal{W}'|}\frac{|\mathcal{K}'|}{|\mathcal{K}|}.$$

Therefore it holds that the probability of generating a weak key when we consider rejection sampling for the first time in our analysis changes by exactly a factor of $(|\mathcal{W}|/|\mathcal{W}'|) \cdot (|\mathcal{K}'|/|\mathcal{K}|)$. This is precisely the probability that a weak key will not be rejected due to weight concerns divided by the probability that key will not be rejected due to weight concerns.

We note that as long as the rejection probabilities for both keys and weak keys is not especially close to 0 or 1, then it is sufficient to sample many keys according to their distributions and observe the portion of these keys that would be rejected.

In order to practically measure the security gained by rejection sampling for the 1-iteration ISD attack against the ephemeral key parameters, we sample 10,000 keys according to $\mathsf{KeyGen'}$ and we sample 10,000 weak keys according to $\mathsf{KeyGen'}$ and we observe how many of them are rejected. We observe that approximately 39.2% of regular keys are rejected while approximately 67.4% of weak keys are rejected. We therefore conclude for this attack and this parameter set, $\frac{|\mathcal{W}|}{|\mathcal{K}|} = 0.582\frac{|\mathcal{W}'|}{|\mathcal{K}'|}$. Therefore, rejection sampling grants less than 1 additional bit of security back to LEDACrypt.

This attack analysis can be efficiently reproduced for additional parameter sets and alternative notions of weak key with the same result.

4.6 Putting it all together

Finally, we re-calculate the results of Section 4.2 using Theorems 4.2 and 4.5, but accounting for the attack improvement of using advanced information set decoding from Section 4.4 and accounting for the security improvement due to rejection sampling issues in Section 4.5. We re-write the formulas with the substitutions of ϵ_{CPA} (resp. ϵ_{CPA}) for the constant $\frac{1}{2}$ for the reader, and note that the definition of ultra-weak keys has been implicitly modified to have more liberal degree constraints to suit the advanced ISD subroutine being used now.

Let x, y, m be defined as in Theorem 4.5. For the case of claimed-256-bit security for ephemeral key parameters, the fraction of ultra-weak keys recovered by a single iteration of the advanced ISD algorithm is

$$d_v{}^2(d_v-1)^2 m \int_{x=0}^{\epsilon}\int_{y=0}^{\epsilon} (xy)^{d_v-2}\left((\epsilon-x)(\epsilon-y)\right)^m \left(\frac{1}{\epsilon-x}+\frac{1}{\epsilon-y}\right) \mathrm{d}y\mathrm{d}x,$$

and the fraction of these ultra-weak keys out of all possible keys is

$$(\epsilon p)d_v{}^2(d_v-1)^2 \int_{x=0}^{\epsilon}\int_{y=0}^{\epsilon} x^{d_v-2}y^{d_v-2}(\epsilon-x)^m(\epsilon-y)^m \left(\frac{m_0{}^2+m_1{}^2}{(\epsilon-x)(\epsilon-y)}+\frac{m_0m_1}{(\epsilon-x)^2}+\frac{m_0m_1}{(\epsilon-y)^2}\right)\mathrm{d}y\mathrm{d}x.$$

Evaluating these formulae with ephemeral key parameters $d_v = 11, m_0 = 7, m_1 = 6, p = 36,877$ and substituting $\epsilon_{\mathsf{CPA}} = .585$ yields 1 key recovered in $2^{62.62}$ per single iteration, and 1 ultra-weak key in $2^{43.90}$ of all possible keys. This yields an algorithm making $2^{62.62-43.90} = 2^{18.72}$ guesses and recovering 1 in $2^{47.72}$ of the ephemeral keys (accounting for the loss due to rejection sampling and the limited number of iterations).

Substituting $\epsilon_{\mathsf{CCA}} = .577$ similarly and evaluating with long-term key parameters $d_v = 13, m_0 = 7, m_1 = 6, p = 152,267$ yields 1 key recovered in $2^{70.45}$ per single iteration and 1 ultra-weak key in $2^{49.55}$ of all possible keys. This yields an algorithm making $2^{70.45-49.55} = 2^{20.90}$ guesses and recovering 1 in $2^{52.54}$ of the long-term keys (accounting for the loss due to rejection sampling and the limited number of iterations).

To conclude, we would like to compare this result against the claimed security level of NIST Category 5. Formally, these schemes should be as hard to break as breaking 256-bit AES. Each guess in the ISD algorithms leads to a cost of approximately p^3 bit operations (due to linear algebra and claw finding operations combined). This is $2^{45.5}$ bit operations for the ephemeral key parameters and $2^{51.6}$ bit operations for the long-term key parameters. A single AES-256 operation costs approximately 2^{15} bit operations. This yields the main result of this section.

Theorem 4.6 (Main). *There is an advanced information set decoding algorithm that costs the same as $2^{49.22}$ AES-256 operations and recovers 1 in $2^{47.72}$ of LEDAcrypt's Category 5 ephemeral keys.*

Similarly, there is an advanced information set decoding algorithm that costs the same as $2^{57.50}$ AES-256 operations and recovers 1 in $2^{52.54}$ of LEDAcrypt's Category 5 long-term keys.

Remark 4. Note that $49.22+47.72 = 96.94 \ll 256$, $57.50+52.54 = 110.03 \ll 256$.

Remark 5. Finally, we recall that we used various heuristics to approximate the above numbers, concretely. However, these simplifying choices can only affect at most one or two bits of security compared to a fully formalized calculation (which would come at the expense of making the analysis significantly more burdensome to parse for the reader).

5 Attack on All Keys

To conclude, we briefly analyze the asymptotic complexity of our new attack strategy in the context of recovering keys in the average case. We first note that, assuming the LEDAcrypt approach is parameterized in a balanced way – that is, H and Q are similarly sparse, and further assuming that n_0 is a constant – the ordinary ISD attack (with a randomly chosen information set) has a complexity of $\exp(\tilde{O}(p^{\frac{1}{2}}))$. To see this, observe that all known ISD variants using a random information set to find an asymptotically sparse secret parity check matrix constructed like the LEDAcrypt private key, have complexity $O\left(\frac{n_0}{n_0-1}\right)^w$, where $w = n_0 d_v m$ is the row weight of the secret parity check matrix. Efficient decoding requires $w = O(p^{\frac{1}{2}})$. By inspection this complexity is $\exp(\tilde{O}(p^{\frac{1}{2}}))$

However, we obtain an improved asymptotic complexity when using structured information sets as follows.

Theorem 5.1. *The asymptotic complexity of ISD using an appropriate choice of structured information sets, when attacking* all *LEDAcrypt keys in the worst case, is* $\exp(\tilde{O}(p^{\frac{1}{4}}))$.

Proof. We analyze the situation with structured information sets. Imagine we are selecting the nonzero coefficients of H' and Q' completely at random, aside from a sparsity constraint. The sparsity constraint needs to be set in such a way that the row weight of the product $H'Q'$ (restricted to two cyclic blocks) has row weight no more than p. This further constrains the row weight of each cyclic block of H' and Q' to be approximately $\left(\frac{p\ln(2)}{n_0}\right)^{\frac{1}{2}} = O(p^{\frac{1}{2}})$. The probability of success per iteration is then at least $O\left(\left(\frac{\ln(2)}{pn_0}\right)^{\frac{1}{2}\cdot(\sum_{i=0}^{n_0-1} m_i+n_0d_v)}\right)$. With balanced parameters, d_v and the m_i are $O(p^{\frac{1}{4}})$, thus the total complexity is indeed $\exp(\tilde{O}(p^{\frac{1}{4}}))$. Note that when H' and Q' are random aside from the sparsity constraint, the probability that the supports of H' and Q' contain the supports of H and Q respectively does not depend on H and Q, so the structured ISD algorithm is asymptotically better than the unstructured ISD algorithm, even when we ignore weak keys. □

Remark 6. The fact that there exists an asymptotically better attack than standard information set decoding against keys structured like those of LEDAcrypt is not itself particularly surprising. Indeed, the very simple attack that proceeds by enumerating all the possible values of H and Q is also asymptotically

$\exp(\tilde{O}(p^{\frac{1}{4}}))$. However, this simple attack does not affect the concrete parameters presented in the Round 2 submission of LEDAcrypt.

In contrast, we strongly suspect, but have not rigorously proven, that our attack significantly improves on the complexity of standard information set decoding against typical keys randomly chosen for some of the submitted parameter sets of LEDAcrypt. In particular, our estimates suggest that the NIST category 5 parameters with $n_0 = 2$ can be attacked with an appropriately chosen distribution for H' and Q' (e.g. with each polynomial block of H' and Q' chosen to have 5 or 6 consecutive chunks of nonzero coefficients in some ring representation) and that typical keys will be broken at least a few hundred times faster than with ordinary information set decoding.

If it were the case that we were attacking an "analogously-chosen" parameter set for LEDAcrypt targeting higher security levels (512-bit security, 1024-bit security, and so on), we believe a much larger computational advantage would be obtained and (importantly) be very easy to rigorously demonstrate.

6 Conclusion

In this work, we demonstrated a novel, real-world attack against LEDAcrypt – one of 17 remaining 2nd Round candidates for standardization in NIST's Post-Quantum Cryptography competition. The attack involved a customized form of Information Set Decoding, which carefully guesses the information set in a non-uniform manner so as to exploit the unique product structure of the keys in LEDAcrypt's design. The attack was most effective against classes of weak keys in the proposed parameter sets asserted to have 256-bit security (demonstrating a trade-off between computational time and fraction of the key space recovered that was better than expected even of a 128-bit secure cryptosystem), but the attack also substantially reduced security of all parameter sets similarly.

Moreover, we demonstrated that these type of weak keys are present throughout the key space of LEDAcrypt, so that simple "patches" such as rejection sampling cannot repair the problem. This was done by demonstrating a continuum of progressively larger classes of less weak keys and by showing that the same style of attack reduces the average-case complexity of certain parameter sets.

Acknowledgements. We thank Corbin McNeill for his contributions to our analysis of rejection sampling on the weak key attack. We also thank the anonymous reviewers for very thorough editorial feedback. This work is funded in part by NSF grant award number 1906360.

References

1. National Institute of Standards and Technology: Post-quantum cryptography project, 2016. `https://csrc.nist.gov/projects/post-quantum-cryptography`.
2. Gorjan Alagic, Jacob Alperin-Sheriff, Daniel Apon, David Cooper, Quynh Quynh Dang, Yi-Kai Liu, Carl Miller, Dustin Moody, Rene Peralta, Ray Perlner, Angela

Robinson, and Daniel Smith-Tone. Status Report on the First Round of the NIST Post-Quantum Cryptography Standardization Process. 2019.

3. Daniel Apon, Ray Perlner, Angela Robinson, and Paolo Santini. Cryptanalysis of LEDAcrypt. Cryptology ePrint Archive, Report 2020/455, 2020. `https://eprint.iacr.org/2020/455`.
4. Anja Becker, Antoine Joux, Alexander May, and Alexander Meurer. Decoding Random Binary Linear Codes in $2n/20$: How $1+1=0$ improves information set decoding. In David Pointcheval and Thomas Johansson, editors, *Advances in Cryptology – EUROCRYPT 2012*, pages 520–536, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.
5. Elwyn Berlekamp, Robert McEliece, and Henk Van Tilborg. On the inherent intractability of certain coding problems (corresp.). *IEEE Transactions on Information Theory*, 24(3):384–386, 1978.
6. Daniel J. Bernstein. Grover vs. McEliece. In Nicolas Sendrier, editor, *Post-Quantum Cryptography*, pages 73–80, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.
7. Daniel J. Bernstein, Tanja Lange, and Christiane Peters. Smaller Decoding Exponents: Ball-Collision Decoding. In Phillip Rogaway, editor, *Advances in Cryptology – CRYPTO 2011*, pages 743–760, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
8. Carlos Aguilar Melchor, Nicolas Aragon, Slim Bettaieb, Loïc Bidoux, Olivier Blazy, Jean-Christophe Deneuville, Philippe Gaborit, Edoardo Persichetti, Gilles Zémor, Jurjen Bos. HQC. available at `https://csrc.nist.gove/projects/post-quantum-cryptography/round-2-submission`, 2019. Technical report, National Institute of Standards and Technology.
9. Daniel J. Bernstein, Tung Chou, Tanja Lange, Ingo von Maurich, Rafael Misoczki, Ruben Niederhagen, Edoardo Persichetti, Christiane Peters, Peter Schwabe, Nicolas Sendrier, Jakub Szefer, Wen Wang. Classic McEliece. available at `https://csrc.nist.gove/projects/post-quantum-cryptography/round-2-submission`, 2019. Technical report, National Institute of Standards and Technology.
10. Lov K. Grover. A fast quantum mechanical algorithm for database search. In *Proc. 28th Annual ACM Symposium on the Theory of Computing*, pages 212–219, Philadephia, PA, May 1996.
11. Ghazal Kachigar and Jean-Pierre Tillich. Quantum Information Set Decoding Algorithms. pages 69–89, 03 2017.
12. Pil Joong Lee and Ernest F. Brickell. An observation on the security of McEliece's public-key cryptosystem. In *Advances in Cryptology - EUROCRYPT 88*, volume 330, pages 275–280. Springer Verlag, 1988.
13. Jeffrey S. Leon. A probabilistic algorithm for computing minimum weights of large error-correcting codes. *IEEE Trans. Information Theory*, 34(5):1354–1359, September 1988.
14. Carl Löndahl, Thomas Johansson, Masoumeh Koochak Shooshtari, Mahmoud Ahmadian-Attari, and Mohammad Reza Aref. Squaring Attacks on McEliece Public-Key Cryptosystems Using Quasi-Cyclic Codes of Even Dimension. *Des. Codes Cryptography*, 80(2):359–377, August 2016.
15. Baldi M., Bodrato M., and Chiaraluce F. A New Analysis of the McEliece Cryptosystem Based on QC-LDPC Codes. *Ostrovsky R., De Prisco R., Visconti I. (eds) Security and Cryptography for Networks. SCN 2008.*, volume 5229, 2008.

16. Marco Baldi, Alessandro Barenghi, Franco Chiaraluce, Gerardo Pelosi, and Paolo Santini. LEDAcrypt. available at `https://csrc.nist.gov/projects/post-quantum-cryptography/round-2-submissions`, 2019. Technical report, National Institute of Standards and Technology.
17. Martin Albrecht, Carlos Cid, Kenneth G. Paterson, Cen Jung Tjhai, Martin Tomlinson. NTS-KEM. available at `https://csrc.nist.gove/projects/post-quantum-cryptography/round-2-submission`, 2019. Technical report, National Institute of Standards and Technology.
18. Alexander May, Alexander Meurer, and Enrico Thomae. Decoding Random Linear Codes in $O(2^{0.054n})$. In Dong Hoon Lee and Xiaoyun Wang, editors, *Advances in Cryptology – ASIACRYPT 2011*, pages 107–124, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.
19. Alexander May and Ilya Ozerov. On Computing Nearest Neighbors with Applications to Decoding of Binary Linear Codes. In Elisabeth Oswald and Marc Fischlin, editors, *Advances in Cryptology – EUROCRYPT 2015*, pages 203–228, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.
20. Robert McEliece. A Public-Key Cryptosystem Based on Algebraic Coding Theory. *The Deep Space Network (DSN) Progress Report*, 44:114–116, 1978.
21. Dustin Moody and Ray Perlner. Vulnerabilities of "McEliece in the World of Escher". In Tsuyoshi Takagi, editor, *Post-Quantum Cryptography*, pages 104–117, Cham, 2016. Springer International Publishing.
22. Nicolas Aragon, Paulo Barreto, Slim Bettaieb, Loic Bidoux, Olivier Blazy, Jean-Christophe Deneuville, Phillipe Gaborit, Shay Gueron, Tim Guneysu, Carlos Aguilar Melchor, Rafael Misoczki, Edoardo Persichetti, Nicolas Sendrier, Jean-Pierre Tillich, Gilles Zemor. BIKE. available at `https://csrc.nist.gove/projects/post-quantum-cryptography/round-2-submission`, 2019. Technical report, National Institute of Standards and Technology.
23. Harald Niederreiter. Knapsack-type cryptosystems and algebraic coding theory. *Prob. Control and Inf. Theory*, 15(2):159–166, 1986.
24. Ray Perlner. Optimizing Information Set Decoding Algorithms to Attack Cyclosymmetric MDPC Codes. In Michele Mosca, editor, *Post-Quantum Cryptography*, pages 220–228, Cham, 2014. Springer International Publishing.
25. Eugene Prange. The use of information sets in decoding cyclic codes. *IRE Transactions on Information Theory*, 8(5):5–9, 1962.
26. Nicolas Sendrier. Decoding One Out of Many. In Bo-Yin Yang, editor, *Post-Quantum Cryptography*, volume 7071 of *Lecture Notes in Computer Science*, pages 51–67. Springer Verlag, 2011.
27. Peter W. Shor. Algorithms for quantum computation: discrete logarithms and factoring. *Proceedings 35th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 124–134, 1994.
28. Jacques Stern. A method for finding codewords of small weight. In *Coding Theory and Applications, 3rd International Colloquium, Toulon, France, November 2-4, 1988, Proceedings*, pages 106–113, 1988.

EXPPO: EXecution Performance Profiling and Optimization for CPS Co-simulation-as-a-Service

Yogesh D. Barve*, Himanshu Neema*, Zhuangwei Kang*, Hongyang Sun*, Aniruddha Gokhale*, Thomas Roth†
*Department of Electrical Engineering and Computer Science
Vanderbilt University
Nashville, TN 37212, USA
{yogesh.d.barve, himanshu.neema, zhuangwei.kang, hongyang.sun, a.gokhale}@vanderbilt.edu
†Smart Grid and Cyber-Physical Systems Program Office, Engineering Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
thomas.roth@nist.gov

***Abstract*—A co-simulation may comprise several heterogeneous federates with diverse spatial and temporal execution characteristics. In an iterative time-stepped simulation, a federation exhibits the Bulk Synchronous Parallel (BSP) computation paradigm in which all federates perform local operations and synchronize with their peers before proceeding to the next round of computation. In this context, the lowest performing (i.e., slowest) federate dictates the progression of the federation logical time. One challenge in co-simulation is performance profiling for individual federates and entire federations. The computational resource assignment to the federates can have a large impact on federation performance. Furthermore, a federation may comprise federates located on different physical machines as is the case for cloud and edge computing environments. As such, distributed profiling and resource assignment to the federation is a major challenge for operationalizing the co-simulation execution at scale. This paper presents the Execution Performance Profiling and Optimization (EXPPO) methodology, which addresses these challenges by using execution performance profiling at each simulation execution step and for every federate in a federation. EXPPO uses profiling to learn performance models for each federate, and uses these models in its federation resource recommendation tool to solve an optimization problem that improves the execution performance of the co-simulation. Using an experimental testbed, the efficacy of EXPPO is validated to show the benefits of performance profiling and resource assignment in improving the execution runtimes of co-simulations while also minimizing the execution cost.**

***Index Terms*—cyber-physical systems, distributed simulation, cloud computing, latency, performance, resource management, gang scheduling**

I. Introduction

Cyber-physical systems (CPS) such as smart city, smart manufacturing, and transactive energy systems must make time-sensitive control decisions to ensure their safe operations. However, such CPS are an amalgamation of multiple dynamic systems such as transportation systems, vehicle dynamics, control systems, and power systems with different interconnected networks of different scales and properties. The assurance that such complex systems are safe and trustworthy requires simulation capabilities that can rapidly integrate tools from multiple domains in different configurations. Co-simulation is an attractive option for interlinking such multiple simulators to simulate higher-level, complex system behaviors.

The IEEE 1516-2010 High Level Architecture (HLA) defines the standardized set of services offered to a process in a distributed co-simulation [1]. A co-simulation in HLA comprises individual simulators called federates that are grouped into a logical entity called a federation. Each federate can have diverse computation and networking resource requirements, which must be considered when assigning resources to the simulators when the federation is deployed to cloud-fog-edge computation environments. The wall-clock execution time required for each federate's computation step varies based on this resource allocation. The variation in execution times can result in some federates waiting on others before they can proceed to the next stage of computation. Federates that require more wall-clock time for their computation steps (the low performing federates) increase the overall completion time (or *makespan*) of the entire simulation. This can potentially violate the simulation completion deadline. Thus, resource allocation and resource configuration selections are very important for the overall performance of distributed simulations.

Although co-simulations have traditionally been hosted in high-performance computing (HPC) clusters, there has been an increasing trend towards the adoption of cloud computing for simulation jobs. It is in this context that the Docker container run-time platform [2] provides a solution for running federates across different computation platforms by providing a unified packaging of simulation code with its software dependencies. But recent research [3] has shown that there are operational challenges, such as performance aware resource assignments, that need to be considered for running simulations in cloud environments. Furthermore, there are several infrastructure-related complexities that must be considered to run these distributed simulations in a cloud environment.

This work presents a performance profiling, simulation run-time optimization and resource configuration platform called EXPPO - (***EX**ecution **P**erformance **P**rofiling and **O**ptimization*). EXPPO uses distributed tracing [4] to assess federate level performance for each computational step. These performance characteristics are used at runtime to determine whether changes to the resource allocation of the federation

Barve, Yogesh; Neema, Himanshu; Kang, Zhuangwei; Sun, Hongyang; Gokhale, Aniruddha; Roth, Thomas. "EXPPO: EXecution Performance Profiling and Optimization for CPS Co-simulation-as-a-Service." Paper presented at the 23rd International Symposium on Real-Time Distributed Computing, Nashville, TN, US. May 19, 2020 - May 21, 2020.

could enable shorter makespan for the simulation run. To enable distributed tracing, EXPPO utilizes the *opentracing* [5] specification for measuring the wall-clock time spent in each computational step of the simulation. To shield the developers from complexities when embedding tracing code in the simulation application logic, EXPPO leverages generative aspect of Model Driven Engineering (MDE) to auto-generate source-code snippets and configuration files for the simulation run. To provide resource recommendations for the individual federates, EXPPO uses the tracing information to build a resource-performance model and solves an optimization problem to find the resource allocation with the lowest makespan and cost for the co-simulation.

The main contributions of EXPPO are:

1) Demonstration that the default resource configuration for a federation has scenarios where a federate with a longer wall-clock execution time for its computational step increases the makespan of the co-simulation;
2) Development of an approach to generate tracing probes inside the source code of the simulation logic, which is required to perform distributed simulation profiling;
3) Development of a new resource recommendation engine which uses an optimization algorithm for finding the resource configuration for a federation that minimizes its overall makespan and cost; and
4) Validation of EXPPO showing its benefits on the performance of distributed simulation execution.

The rest of the paper is organized as follows. Section II provides a motivating use-case for EXPPO and lists its key requirements. Section III provides an overview of the EXPPO framework. Section IV describes the key EXPPO components and its resource configuration and optimization algorithms. EXPPO 's cloud architecture and co-simulation framework are described in Section V. The experiment evaluation results are described in Section VI, related works are described in Section VII and Section VIII concludes the paper.

II. Motivation and Solution Requirements

The computation pattern of many time-stepped co-simulations follows the Bulk Synchronous Parallel (BSP) model [6]. In this model, every participating simulation completes its computation for a given time step, waits for the other participating simulations to complete their computations, exchanges new state information with its peers, and only then proceeds to the next time step. Thus, if one simulation takes more time to execute, the other simulations are forced to wait for it and remain idle. This is illustrated in Figure 1 which shows an HLA federation with three federates. Each federate executes one computation task repeated each time step. Federate 1 takes the longest to execute, and the other two federates are waiting for Federate 1 to complete before moving to the next computation step. This increases the makespan of the entire simulation, thereby decreasing the performance of the simulation execution.

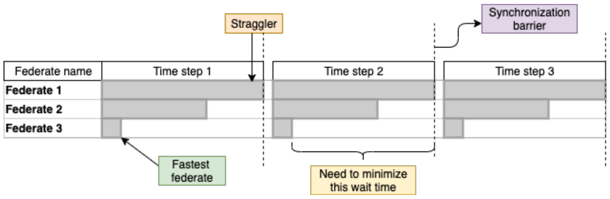


Fig. 1: Performance visualization of an example federation.

Recently, co-simulations have been deployed using container management solutions [7] [8], such as Docker Swarm [9], Kubernetes [10], Mesos [11]. However, these solutions use queue-based scheduling, wherein the containers are allocated to machines one at a time. Hence, there could be instances where there may not be enough resources available in the cluster to schedule all the federates of the federation. However, a few federates may still get deployed while the remaining federates are stuck in the scheduler queue until more resources are available. This will cause the entire simulation to stall, since all the federates are required to participate in the simulation to progress. For instance, from Figure 1, if there are resources to deploy Federate 1 and Federate 3, and not sufficient resources to deploy Federate 2, the job scheduler should not deploy any federates and should wait until more resources are available. Hence, there is a need for bag-of-tasks scheduling mechanism for deploying these federates on the cloud computing environments.

The resource configuration also plays an important role in the execution time of the federates. Figure 2(a) depicts the performance of a federate when assigned different numbers of cores for executing a computation step. This federate is running a Freqmine application from the Princeton Application Repository for Shared-Memory Computers (PARSEC) benchmark [12] as its computation task. This figure shows that the execution time generally decreases with the increasing amount of resources assigned to the federate. Thus, there is a potential for minimizing the wait time by appropriately configuring the resources assigned to the federates. Minimizing the wait time can be done either by providing the highest resource configuration for all the federates, or finding a resource configuration that considers the cost of assigning the resources to the federates. However, deciding what resource configuration to select for a given computation is a non-trivial task for a simulation developer who may not have the domain expertise of configuring and running applications in cloud computing environments. Performance profiling of the federates can help in understanding the relation between the resource assignment and the execution performance. However, these simulations might be deployed across different physical and/or virtual host environments when running in cloud computing platforms; performance profiling for such distributed simulations can be very challenging.

Building on the above use-case, below are the four key requirements EXPPO aims to satisfy:

- **Requirement R1.** *Conduct distributed performance tracing of the federates:* To understand the bottlenecks in federation performance, there is a need for logging and

Barve, Yogesh; Neema, Himanshu; Kang, Zhuangwei; Sun, Hongyang; Gokhale, Aniruddha; Roth, Thomas. "EXPPO: EXecution Performance Profiling and Optimization for CPS Co-simulation-as-a-Service." Paper presented at the 23rd International Symposium on Real-Time Distributed Computing, Nashville, TN, US. May 19, 2020 - May 21, 2020.

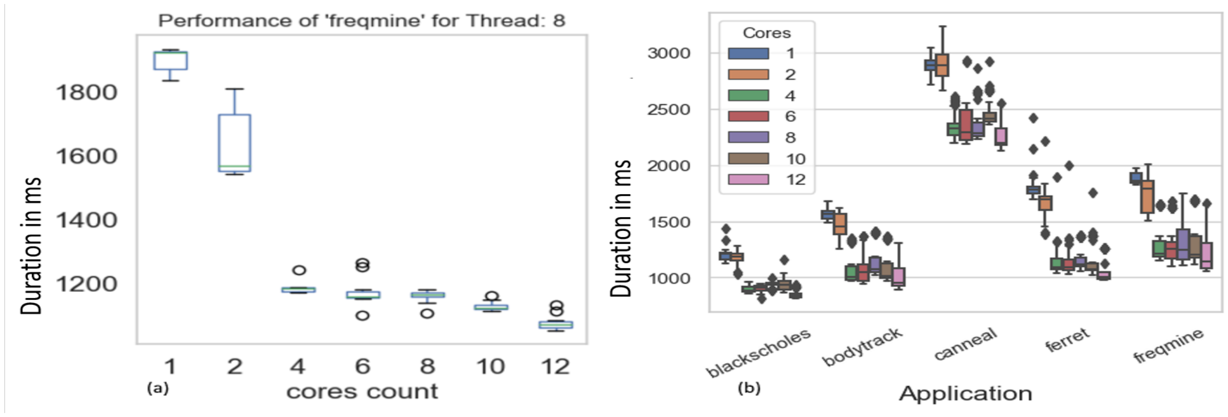


Fig. 2: (a) Performance of the federate running the PARSEC Freqmine application using 8 threads for different resource configuration selections. (b) Performance of five PARSEC benchmark applications for different resource configuration selections.

gathering execution performance traces of the federates. The tracing infrastructure needs to handle federates which are distributed across multiple physical hosts. Hence, the tracing infrastructure should be able to correlate traces from different federates of a federation.

- **Requirement R2.** *Reduce complexity in provisioning software probes:* Requiring the developer to manually write source code for performance tracing for the federation can result in complexities and errors which need to be minimized. The developer needs to understand the tracing software and write code which adheres to the tracing software requirements. This is tedious for the developer, who now apart from writing the simulation logic must also setup and configure the tracing infrastructure. EXPPO should reduce the manual configuration of the tracing information, thereby reducing repeated effort by the developer.
- **Requirement R3.** *Recommend resource configuration to minimize the co-simulation makespan and cost:* It can be challenging for an end user to determine the resource requirements for a federation because each federate can be configured differently. A bad resource configuration can have inadvertent effect on the completion time of the simulation. However, the choice of resource configuration also has an associated cost. Hence, the configuration must be chosen such that it satisfies both quality of service (QoS) and cost factors.
- **Requirement R4.** *Provide a gang-scheduling algorithm for executing simulations:* The runtime platform should support deployment of multiple federate using bag-of-tasks scheduling (or gang scheduling) algorithm.

III. Overview of EXPPO

Figure 3(a) shows the workflow and components of EXPPO. The *design phase* requires the developer to model the federation using the federation development toolkit. This toolkit is based on the Web-based Generic Modeling Environment (WebGME) [13]. To measure the execution time of a federate in a time step, the simulator needs to embed a tracing code initializer and logger to record the execution time completion of the computation step. The tracing initialization code and the tracer configuration files are auto-generated by the custom WebGME model interpreters. Once the required code is generated and the user has implemented the necessary simulation logic, the federate is compiled into an executable image, which is then packaged inside a Docker container.

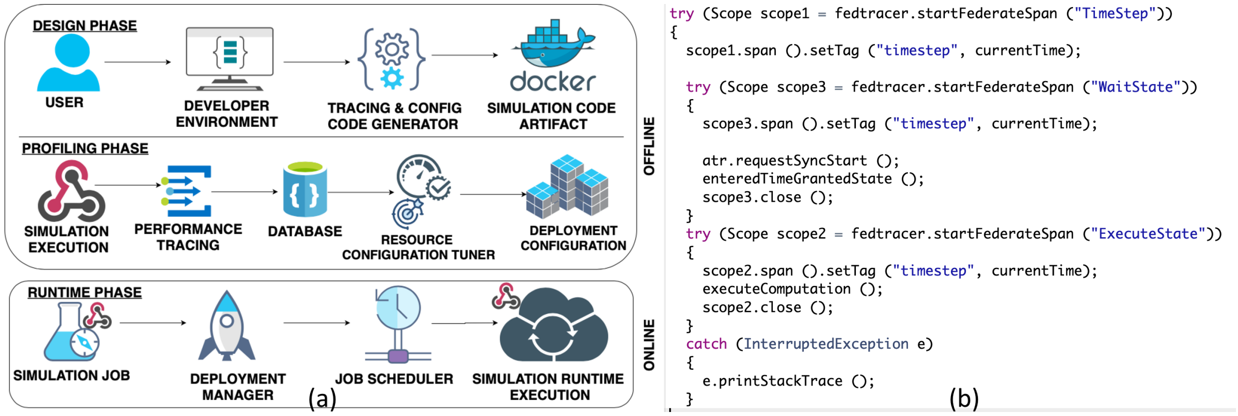
During the *profiling phase*, each federate is executed and profiled under different resource configurations. The execution time for each federate is logged using the tracing information, and this tracing information is stored in a centralized database. The resource configuration tuner uses the recorded logs together with user provided objectives to optimize the resource configuration for each federate. Finally, the optimized resource configuration is used to configure the co-simulation deployment accordingly.

During the *runtime phase*, the simulation job information for the federates in the co-simulation is submitted to the deployment manager. The job information includes the name of the federate Docker image, resource configuration requirements, etc. The deployment manager submits the scheduling information to the job scheduler, which handles the execution of the jobs on the co-simulation runtime execution platform.

IV. Design Elements of EXPPO

EXPPO allows users to design and deploy co-simulations on distributed compute infrastructures supported by Docker-based virtualization. Its generative capabilities simplify the auto-generation of performance monitoring instrumentation, configuration and probing for the different federates. Its resource configuration tuner optimizes resource allocations to federates to lower the makespan and execution cost for the federation execution. Its runtime platform supports parallel execution of different federations on its shared compute infrastructure. This section details each of the EXPPO components.

Barve, Yogesh; Neema, Himanshu; Kang, Zhuangwei; Sun, Hongyang; Gokhale, Aniruddha; Roth, Thomas. "EXPPO: EXecution Performance Profiling and Optimization for CPS Co-simulation-as-a-Service." Paper presented at the 23rd International Symposium on Real-Time Distributed Computing, Nashville, TN, US. May 19, 2020 - May 21, 2020.



```
try (Scope scope1 = fedtracer.startFederateSpan ("TimeStep"))
{
  scope1.span ().setTag ("timestep", currentTime);

  try (Scope scope3 = fedtracer.startFederateSpan ("WaitState"))
  {
    scope3.span ().setTag ("timestep", currentTime);

    atr.requestSyncStart ();
    enteredTimeGrantedState ();
    scope3.close ();
  }
  try (Scope scope2 = fedtracer.startFederateSpan ("ExecuteState"))
  {
    scope2.span ().setTag ("timestep", currentTime);
    executeComputation ();
    scope2.close ();
  }
  catch (InterruptedException e)
  {
    e.printStackTrace ();
  }
```

Fig. 3: (a) Workflow of EXPPO illustrating the connections between different components of the system. (b) Profiling code snippet in Java language generated leveraging the MDE techniques.

A. *Performance Profiling of Federates*

Understanding the performance characteristics of the co-simulation execution is of paramount importance when deciding how to run federates in runtime execution environments. Individual federates have different resource needs, and their execution performance will vary depending on how the resources are assigned. Thus, understanding the performance profiles of each federate is critical to the problem of optimizing the resource allocation and thereby lowering the makespan and execution cost of the federation execution. EXPPO leverages distributed tracing to assist in the logging of time stamps of distributed events generated in the federation (Requirement **R1**). It leverages Opentracing instrumentation [5] to track execution time spent during each computation time step of the federate. The execution time is then logged into a timeseries database for conducting performance analysis. It leverages MDE technologies [14] such as Domain Specific Modeling Language (DSML), code generators and model interpreters to synthesize performance profiling software artifacts which can be used for performance profiling of federates (Requirement **R2**). An example code snippet is shown in Figure 3(b).

B. *Federation Resource Configuration Optimization*

When running a federate in a Docker container, cloud providers usually have multiple resource configurations from which the user can select for their application. However, without analyzing the resource dependency of the application, it may be challenging for the user to select the resource configuration that meets the application's quality of service (QoS) requirement while at the same time minimizing the execution cost in terms of the cost of resources utilized. Figure 2(b) shows how the different resource configuration impacts the execution time of the federate which is running applications from the PARSEC benchmark. Furthermore, in a co-simulation, the resource selection becomes critical as every federate's execution time will contribute to the co-simulation's performance. To address this issue (Requirement **R3**), EXPPO provides a resource configuration recommendation system that selects the resource assignment which optimizes the application's QoS performance and user's budget requirements.

Consider the following optimization problem: Suppose the system has a set of homogeneous machines, each with k processing cores. Let $F = \{f_1, f_2, \ldots, f_n\}$ denote a federation (co-simulation) that consists of a set of n federates. Given a resource assignment $R = [r_1, r_2, \ldots, r_n]$ to each federate, the execution time of federate $f_j \in F$ can be expressed as $t_j(r_j)$ when assigned r_j cores. For performance reasons, assume a federate cannot be split among two or more machines, so we have $1 \leq r_j \leq k$. The makespan M for every computation step for the entire federation F is dictated by the slowest running federate (i.e., the straggler), and is defined as $M = \max_j t_j(r_j)$. The execution cost C is given by the total resource used by all the federates over the makespan duration. Since the cores will be reserved for the federation until the slowest federate is done executing, the cost is defined as $C = M \cdot \sum_j r_j$. The resource configuration recommender needs to find a resource assignment R^* that minimizes $G = \alpha M + \beta C = M(\alpha + \beta \sum_j r_j)$, where α and β denote the user-defined weights to the application's QoS and the execution cost, respectively.

Two additional assumptions are made to solve this optimization problem: (1) federate execution time does not increase with the amount of resources (number of cores) assigned, i.e., $r_j \leq r'_j$ implies $t_j(r_j) \geq t_j(r'_j)$; (2) federate execution cost does not decrease with the amount of resources assigned, i.e., $r_j \leq r'_j$ implies $c_j(r_j) \leq c_j(r'_j)$, where $c_j(r) = r \cdot t_j(r)$. These are realistic assumptions as many practical applications have *monotonically increasing and sublinear* speedup functions [15], [16], such as those that follow Amdahl's law [17]. As such, the optimization problem can be solved by examining all possible makespan values while guaranteeing the minimum cost. Algorithm 1 presents the pseudocode of this solution with a time complexity of $O(n \log n)$ by maintaining a priority queue for all jobs. Similar approaches can be applied to

Algorithm 1: Resource Configuration Tuner

```
Input  : Execution time $t_j(r)$ for each federate $f_j$ in
         federation $F$ when allocated different amounts of
         resources $r$, where $1 \le r \le k$.
Output: A resource assignment $R^* = [r_1, r_2, \dots, r_n]$ for
         each federate in the federation that minimizes a
         linear combination of makespan and cost.
Initialize $r_j \leftarrow 1$ for all $1 \le j \le n$;
Compute $G \leftarrow (\max_j t_j(r_j)) \cdot (\alpha + \beta \sum_j r_j)$;
$R^* \leftarrow [r_1, r_2, \dots, r_n]$ and $G^* \leftarrow G$;
while $\sum_j r_j < nk$ do
    $j \leftarrow$ Index of a federate with longest execution time;
    if $r_j = k$ then
        break;
    else
        Increment $r_j$ to the next higher profiled resource
         amount;
        Update $G \leftarrow (\max_j t_j(r_j)) \cdot (\alpha + \beta \sum_j r_j)$;
        if $G < G^*$ then
            $R^* \leftarrow [r_1, r_2, \dots, r_n]$ and $G^* \leftarrow G$;
```

find the minimum makespan subject to a cost budget or the minimum cost for a target makespan.

C. Federation Machine Scheduling Heuristics

A custom scheduler is required to deploy all the federates in the federation to their respective distributed computing environments. Since the federates cannot run independently of the federation, the scheduling scheme must simultaneously run all of the federates of the federation (Requirement **R4**). This is referred to as *Gang scheduling* or *Bag-of-tasks scheduling* in the literature. To achieve this, EXPPO supports two heuristics to simultaneously schedule the federates on a fixed number m of available machines while utilizing the resource configuration results obtained from Section IV-B. These approaches handle the case where some of the machines are loaded with other compute tasks unrelated to the federation.

The first heuristic is inspired by the First-Fit Decreasing (FFD) algorithm for bin packing and is described in Algorithm 2. The heuristic first sorts all the federates in decreasing order of resource assignment and then tentatively allocates each one of them in order onto the first available machine. If all federates in the federation can be successfully allocated, then the schedule is finalized; otherwise, the entire federation will be temporarily put in a waiting queue to be scheduled later. The time complexity of the heuristic is $O(n(\log n + m))$. The other heuristic is based on the Best-Fit Decreasing (BFD) algorithm that works similarly to FFD, except that it finds, for each federate, a best-fitting machine (i.e., with the least remaining resource after hosting the federate). Note that since finding the optimal schedule (or bin packing) is an NP-complete problem, these heuristics may not always find a feasible allocation for a federation even if one exists. However, once an allocation has been found, it is guaranteed to produce the optimal makespan and cost for the federation by using the resource configuration from Section IV-B.

Algorithm 2: First Fit Decreasing (FFD)

```
Input  : Resource assignment $R = [r_1, r_2, ..., r_n]$ for all
         federates in a federation. Current available
         resource $A = [a_1, a_2, ..., a_m]$ of all $m$ machines
         in the system.
Output: Machine allocation $L = [\ell_1, \ell_2, \dots, \ell_n]$ of all
         federates in the system.
Sort all resource assignments in decreasing order, i.e.,
  $r_1 \ge r_2 \ge \cdots \ge r_n$;
Initialize $\ell_j \leftarrow 0, \ \forall 1 \le j \le n$;
for $j = 1, 2, \dots, n$ do
    $fit \leftarrow false$;
    for $i = 1, 2, \dots, m$ do
        if $r_j \le a_i$ then
            Update $a_i \leftarrow a_i - r_j$;
            Set $\ell_j \leftarrow i$;
            $fit \leftarrow true$;
            break;
    if $fit = false$ then
        // revert allocations done so far
        for $k = 1, 2, \dots, j-1$ do
            $i \leftarrow \ell_k$;
            $a_i \leftarrow a_i + r_k$;
            $\ell_k \leftarrow 0$;
        break;
```

V. Co-simulation as a Service Middleware

Figure 4 shows the different components and workflow of the EXPPO co-simulation framework. It is called the Co-simulation-as-a-Service (CaaS) middleware because it enables users to automatically deploy groups of service-based applications to a cloud environment without any concern of resource allocation, application lifecycle monitoring, and cluster management. The functionality of each component is described below.

In ①, the FrontEnd component allows a user to submit a simulation job descriptor in JavaScript Object Notation (JSON) using a Representational State Transfer (REST) Application Programming Interface (API). Each simulation job contains a list of federates, and each federate is run on an individual Docker container. The simulation job descriptor also includes meta-information for each federate, for example, resources required, running status, container image details, etc. The FrontEnd creates a record in a database ④ for each incoming job, then relays the job identifier to JobManager ② which handles resource management. Instead of deploying jobs immediately, the JobManager stores received jobs in a local queue and consults the database about the latest status of cluster resources. It then periodically transfers the information of pending jobs and available resources to JobScheduler ③, which is a pluggable component that implements multiple scheduling algorithms. The JobScheduler either replies with a scheduling decision if the submitted jobs are deployable, or returns a KeepWaiting signal to notify the JobManager that resources are insufficient. The JobManager then forwards deployable jobs to GlobalManager ⑤, synchronizes job status

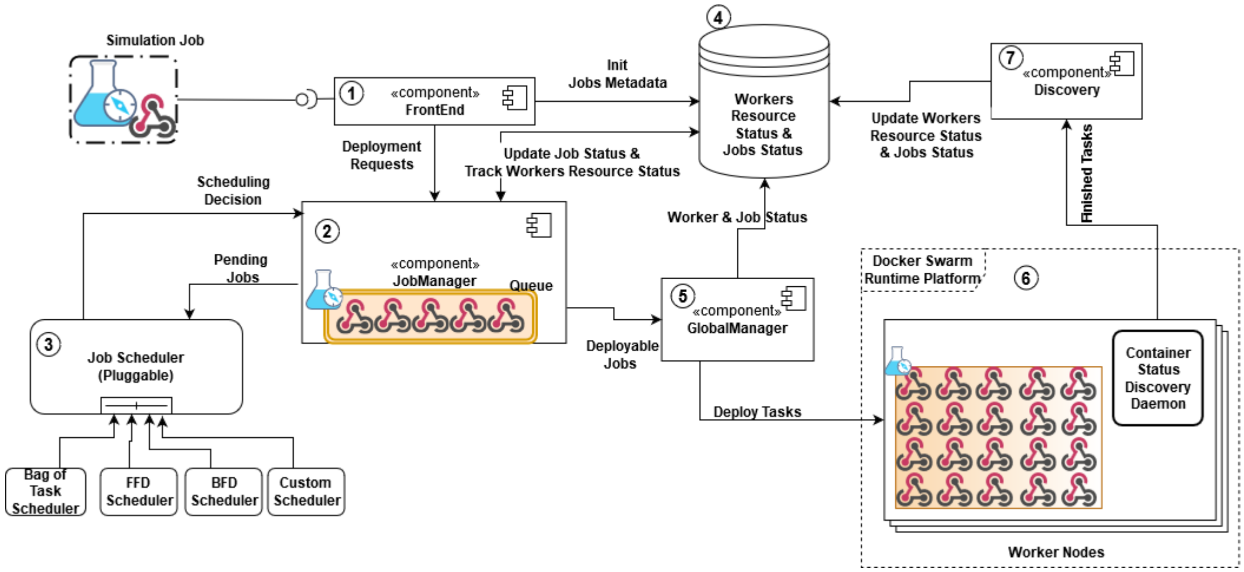


Fig. 4: Co-simulation-as-a-Service.

with the database, and deletes the deployed jobs from its queue. The GlobalManager is responsible for managing participants of the Docker Swarm Runtime Platform ⑥. It launches the master node of the Docker Swarm cluster and accepts registration requests sent from worker nodes. Every joined Worker Node runs a CaaS-Worker daemon that is used to receive and perform commands sent from the GlobalManager. The GlobalManager parses received deployment requests and spawns containers in specific Worker Nodes. Additionally, the Worker Nodes employ a CaaS-Discovery daemon to track the status of containers, and reports a StatusChanged signal to the Discovery component ⑦ when a task is completed.

VI. Experimental Evaluation

A. Experimental Setup

EXPPO was validated using seven homogeneous compute servers with a configuration of 12-core 2.1 GHz AMD Opteron central processing units, 32 GB memory, 500 GB disk space, and the Ubuntu 16.04 operating system. The runtime platform was based on Docker engine version 19.0.5 with swarm mode enabled. There was one client machine that submitted simulation job requests. The front end, job manager, job scheduler and the global manager components were running on a single shared compute server, and five compute worker servers were deployed for running the simulation jobs.

The simulation job consisted of three federates each running a unique application from the PARSEC benchmark: freqmine, blackscholes and ferret. These applications were chosen as they are realistic representations of real-world simulation tasks [12]. During the federation execution, each federate executed its application at every logical time step, for a total number of 100 logical time steps. The implementation of the co-simulation federation was done using Portico HLA [18]. The BFD scheduler was used in the experiments (as it was found to have better performance than FFD). The weights for the QoS and the cost were set to $\alpha = 1$ and $\beta = 0.5$. These weights are chosen so that they correspond to a specific brand of QoS that prioritizes low makespan over resource usage, which is representative of CPS applications.

B. Experimental Results

EXPPO recommended a resource configuration of 4 cores for freqmine federate, 4 cores for ferret federate and 1 core for blackscholes federate. Two baseline approaches were used to compare the performance of resource configuration selection of EXPPO. In the first approach (*least configuration*), all the federates were assigned the lowest possible configuration of 1 core each. In the second approach (*max configuration*), all the federates were assigned the highest possible configuration of 10 cores each. The performance data was collected over 10 simulation jobs.

Figure 5(a) shows the cumulative distribution function (CDF) of the execution time of EXPPO compared to the other two approaches. The resource configurations selected by EXPPO for the federation had a 90*th* percentile execution time of around 230 seconds, which was significantly better than the 320 seconds for the *least configuration*. Its performance was close to that of the *max configuration* (90*th* percentile execution time of around 200 seconds), with the difference due to its lower resource allocation to conserve the cost.

Figure 5(b) shows the cost analysis of the three strategies using the cost function defined in Section IV-B. As can be seen, resource configuration selected by EXPPO incurred a larger cost than the *min configuration* due to the higher resource allocation to reduce execution time, but it had a substantially lower cost compared to the *max configuration*.

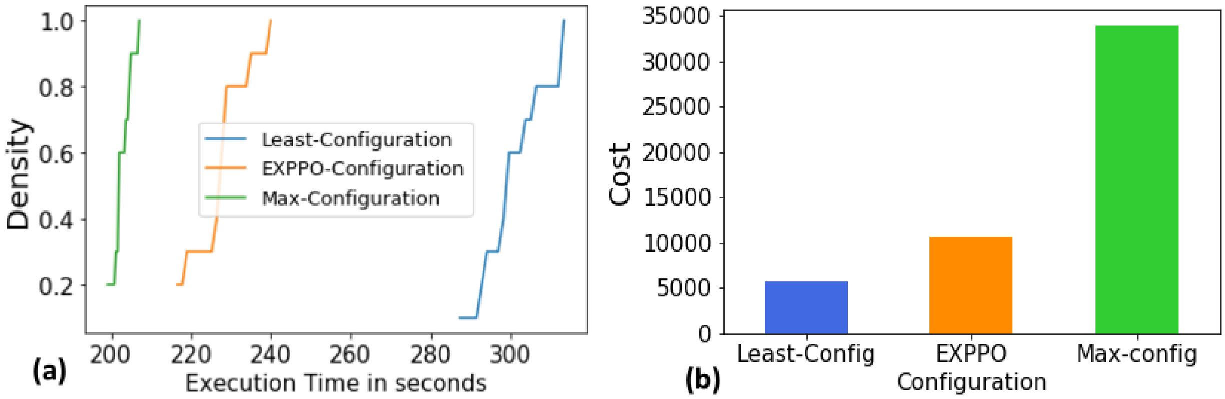


Fig. 5: (a) Execution time for the EXPPO resource configuration compared to other strategies. (b) Cost for the EXPPO resource configuration compared to other strategies.

Overall, the results show that EXPPO is able to select resource configurations and schedule the federates in such a way that minimizes the combination of execution time and cost of the simulation successfully.

VII. Related Work

In [7], the authors presented a Kubernetes co-simulation execution platform for cloud computing environments. Similarly, [8] presented a Docker swarm co-simulation platform for running mixed electrical energy systems simulations. However, these platforms do not use a gang-scheduling based simulation deployment strategy.

For resource recommendation, [19] presented a data-driven approach for selecting the best resource configuration for a virtual machine from a set of different configuration options. [20] explored the cost-sensitive allocation of independent tasks to cloud computing environments. However, these approaches are different from EXPPO as they focus on a single task rather than a pool of BSP tasks.

In [21], the authors proposed a scientific workflows scheduling algorithm to minimize the execution time under budget constraints for deploying to cloud computing environments. [22] proposed an advance-reservation scheduling strategy for message passing interface (MPI) applications. Similarly, [23] presented an approach for gang-scheduling of jobs with different resource needs, such as a compute intensive task paired with a network or I/O intensive task. In [24], the authors present scheduling of multiple container workloads on shared cluster as a minimum cost flow problem (MCFP) constraint satisfaction problem. In [25], the authors proposed a locality-based process placements for parallel and distributed simulation. The evaluation of the proposed framework was carried out using the OMNET++ network simulator.

Compared to related work, EXPPO provides a resource recommendation engine which tries to minimize the makespan and cost for the entire federation (BSP tasks) rather than a single task. EXPPO uses a gang scheduling scheme based on heuristic bin packing techniques to deploy the entire federation on Docker container platform as one batch job. Furthermore, EXPPO provides automatic code generation of distributed tracing probes for performance profiling of the federates which maybe deployed on a distributed infrastructure, thereby relieving the developers from incurring complexities in writing code for the profiling of federates [26].

VIII. Conclusion & Future Works

Resource allocation plays a critical role in co-simulation performance. However, the end user is not necessarily well-equipped to determine what resource allocations work best for their co-simulation jobs given the various resource configuration options (and associated costs) available from the cloud provider. To address these challenges, this paper presents EXPPO, which is a Co-simulation-as-a-Service (CaaS) platform for executing distributed co-simulations in cloud computing environments. EXPPO provides performance profiling capabilities for federates which help in the understanding of the relationship between resource allocations and the simulation performance. Similarly, it addresses performance profiling challenges for a simulation job comprising heterogeneous computation tasks or federates deployed across distributed systems. Furthermore, EXPPO selects the resource configurations for these federates in a way that not only minimizes the makespan of the co-simulation, but also satisfies the cost budget of the user.

Future work will address the following:

- The assumption that federate computations have identical wall-clock execution times across all iterations restricts the generality of the approach to a subset of application use cases. For hybrid simulations or simulations with non-linear dynamics, the execution times of the involved simulation steps will vary at each iteration. Future work must explore dynamic resource allocation for federates which have different work loads for different computation steps. Reinforcement learning approaches which can monitor the federates and dynamically adjust resource allocations based on the varying demand of a federate over time offer a promising approach that remains to be explored.

- EXXPO only considers workloads that are CPU bound. It assumes constant communication costs during each computation step for all the federates. Future work must consider the different communication costs for simulations which may need to be constantly updating states or sending messages for triggering discrete events.
- When other applications are co-located in the cluster environments, the effects of noisy-neighbors [27], [28] can also affect federation performance. Thus, future work could consider the effect of noisy-neighbors for resource scheduling and simulation placement in the cluster.
- EXPPO assumes the use of homogeneous servers for running simulations to simplify its algorithms, and could be extended to include heterogeneous systems. Also, with the growing relevance of edge computing and digital twin techniques, efficient resource allocation and scheduling of the simulations at the edge will be necessary.

IX. Acknowledgments

This work was supported in part by the National Institute of Standards and Technology under award 70NANB18H269; and AFOSR DDDAS FA9550-18-1-0126. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States. Certain commercial products are identified in order to adequately specify the procedure; this does not imply endorsement or recommendation by NIST, nor does it imply that such products are necessarily the best available for the purpose.

References

[1] "IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA)-Framework and Rules." Institute of Electrical and Electronic Engineers New York, 2010, *doi*:10.1109/ieeestd.2000.92296.

[2] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.

[3] S. J. Taylor, A. Khan, K. L. Morse, A. Tolk, L. Yilmaz, J. Zander, and P. J. Mosterman, "Grand challenges for modeling and simulation: simulation everywhere—from cyberinfrastructure to clouds to citizens," *Simulation*, vol. 91, no. 7, pp. 648–665, 2015, *doi*:10.1177/0037549715590594.

[4] J. Mace and R. Fonseca, "Universal context propagation for distributed system instrumentation," in *Proceedings of the Thirteenth EuroSys Conference*. ACM, 2018, p. 8, *doi*:10.1145/3190508.3190526.

[5] Opentracing, "Opentracing specification," https://opentracing.io/specification/, 2020.

[6] T. L. Williams and R. J. Parsons, "The heterogeneous bulk synchronous parallel model," in *International Parallel and Distributed Processing Symposium*. Springer, 2000, pp. 102–108, *doi*:10.1007/3-540-45591-4$_1$2.

[7] K. Rehman, O. Kipouridis, S. Karnouskos, O. Frendo, H. Dickel, J. Lipps, and N. Verzano, "A cloud-based development environment using hla and kubernetes for the co-simulation of a corporate electric vehicle fleet," in *2019 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2019, pp. 47–54, *doi*:10.1109/sii.2019.8700423.

[8] Y. Barve, H. Neema, S. Rees, and J. Sztipanovits, "Towards a design studio for collaborative modeling and co-simulations of mixed electrical energy systems," in *2018 IEEE International Science of Smart City Operations and Platforms Engineering in Partnership with Global City Teams Challenge (SCOPE-GCTC)*. IEEE, 2018, pp. 24–29, *doi*:10.1109/scope-gctc.2018.00011.

[9] Docker, "Docker swarm," https://docs.docker.com/engine/swarm/, 2020.

[10] Kubernetes, "Kubernetes :production-grade container orchestration," https://kubernetes.io/, 2020.

[11] B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A. D. Joseph, R. H. Katz, S. Shenker, and I. Stoica, "Mesos: A platform for fine-grained resource sharing in the data center." in *NSDI*, vol. 11, no. 2011, 2011, pp. 22–22.

[12] C. Bienia, S. Kumar, and K. Li, "Parsec vs. splash-2: A quantitative comparison of two multithreaded benchmark suites on chip-multiprocessors," in *Workload Characterization, 2008. IISWC 2008. IEEE International Symposium on*. IEEE, 2008, pp. 47–56, *doi*:10.1109/iiswc.2008.4636090.

[13] M. Maróti, R. Kereskényi, T. Kecskés, P. Völgyesi, and A. Lédeczi, "Online collaborative environment for designing complex computational systems," *Procedia Computer Science*, vol. 29, pp. 2432–2441, 2014, *doi*:10.1016/j.procs.2014.05.227.

[14] Y. Barve, S. Shekhar, S. Khare, A. Bhattacharjee, and A. Gokhale, "UPSARA: A Model-driven Approach for Performance Analysis of Cloud-hosted Applications," in *11th IEEE/ACM International Conference on Utility and Cloud Computing (UCC)*, Zurich, Switzerland, Dec. 2018, pp. 1–10, *doi*:10.1109/ucc.2018.00009.

[15] M. D. Hill and M. R. Marty, "Amdahl's law in the multicore era," *Computer*, vol. 41, no. 7, pp. 33–38, 2008, *doi*:10.1109/hpca.2008.4658638.

[16] B. Berg, J. L. Dorsman, and M. Harchol-Balter, "Towards optimality in parallel scheduling," *POMACS*, vol. 1, no. 2, pp. 40:1–40:30, 2017, *doi*:10.1145/3154499.

[17] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, ser. AFIPS '67 (Spring), 1967, pp. 483–485, *doi*:10.1145/1465482.1465560.

[18] Portico, "Portico," https://github.com/openlvc/portico, 2016, *doi*:10.1093/acrefore/9780199381135.013.5262.

[19] N. J. Yadwadkar, B. Hariharan, J. E. Gonzalez, B. Smith, and R. H. Katz, "Selecting the best vm across multiple public clouds: A data-driven performance modeling approach," in *Proceedings of the 2017 Symposium on Cloud Computing*. ACM, 2017, pp. 452–465, *doi*:10.1145/3127479.3131614.

[20] S. Selvarani and G. S. Sadhasivam, "Improved cost-based algorithm for task scheduling in cloud computing," in *2010 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2010, pp. 1–5, *doi*:10.1109/iccic.2010.5705847.

[21] X. Lin and C. Q. Wu, "On scientific workflow scheduling in clouds under budget constraint," in *2013 42nd International Conference on Parallel Processing*. IEEE, 2013, pp. 90–99, *doi*:10.1109/icpp.2013.18.

[22] I. Foster, C. Kesselman, C. Lee, B. Lindell, K. Nahrstedt, and A. Roy, "A distributed resource management architecture that supports advance reservations and co-allocation," in *1999 Seventh International Workshop on Quality of Service. IWQoS'99.(Cat. No. 98EX354)*. IEEE, 1999, pp. 27–36, *doi*:10.1109/iwqos.1999.766475.

[23] Y. Wiseman and D. G. Feitelson, "Paired gang scheduling," *IEEE transactions on parallel and distributed systems*, vol. 14, no. 6, pp. 581–592, 2003, *doi*:10.1109/tpds.2003.1206505.

[24] Y. Hu, H. Zhou, C. de Laat, and Z. Zhao, "Concurrent container scheduling on heterogeneous clusters with multi-resource constraints," *Future Generation Computer Systems*, vol. 102, pp. 562–573, 2020, *doi*:10.1016/j.future.2019.08.025.

[25] S. Zaheer, A. W. Malik, A. U. Rahman, and S. A. Khan, "Locality-aware process placement for parallel and distributed simulation in cloud data centers," *The Journal of Supercomputing*, vol. 75, no. 11, pp. 7723–7745, 2019, *doi*:10.1007/s11227-019-02973-9.

[26] A. Bhattacharjee, Y. Barve, A. Gokhale, and T. Kuroda, "(wip) cloudcamp: Automating the deployment and management of cloud services," in *2018 IEEE International Conference on Services Computing (SCC)*. IEEE, 2018, pp. 237–240, *doi*:10.1109/scc.2018.00038.

[27] Y. D. Barve, S. Shekhar, A. Chhokra, S. Khare, A. Bhattacharjee, Z. Kang, H. Sun, and A. Gokhale, "Fecbench: A holistic interference-aware approach for application performance modeling," in *2019 IEEE International Conference on Cloud Engineering (IC2E)*, June 2019, pp. 211–221, *doi*:10.1109/IC2E.2019.00035.

[28] S. Shekhar, Y. Barve, and A. Gokhale, "Understanding performance interference benchmarking and application profiling techniques for cloud-hosted latency-sensitive applications," in *Proceedings of the10th International Conference on Utility and Cloud Computing*. ACM, 2017, pp. 187–188, *doi*:10.1145/3147213.3149453.

3D BUILD MELT POOL PREDICTIVE MODELING FOR POWDER BED FUSION ADDITIVE MANUFACTURING

Zhuo Yang[1,2], Yan Lu *[2], Ho Yeung[2], and Sundar Kirishnamurty[1]

[1]University of Massachusetts, Amherst, MA, USA
[2]National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA

ABSTRACT

Melt pool size is a critical intermediate measure that reflects the outcome of a laser powder bed fusion process setting. Reliable melt pool predictions prior to builds can help users to evaluate potential part defects such as lack of fusion and over melting. This paper develops a layer-wise Neighboring-Effect Modeling (L-NBEM) method to predict melt pool size for 3D builds. The proposed method employs a feedforward neural network model with ten layer-wise and track-wise input variables. An experimental build using a spiral concentrating scan pattern with varying laser power was conducted on the Additive Manufacturing Metrology Testbed at the National Institute of Standards and Technology. Training and validation data were collected from 21 completed layers of the build, with 6,192,495 digital commands and 118,928 in-situ melt pool coaxial images. The L-NBEM model using the neural network approach demonstrates a better performance of average predictive error (12.12%) by leave-one-out cross-validation method, which is lower than the benchmark NBEM model (15.23%), and the traditional power-velocity model (19.41%).

Keywords: Melt pool size, powder bed fusion, additive manufacturing, machine learning, layer-wise, track-wise

1 INTRODUCTION

Laser powder bed fusion (LPBF) additive manufacturing (AM) uses a laser to melt and fuse spread powder on a build plate. An LPBF machine can precisely scan thin layers of powder to form a designed geometry. The laser delivers thermal energy to the powder to create melt pools once it reaches the melting temperature. Under ideal conditions, the laser should remelt the nearby solidified part to bond the newly melted part [1]. This process occurs both horizontally (track-wise) and vertically (layer-wise) [2,3].

Melt pool size is a generalized term that represents a group of values such as depth, width, and length. These measurements closely correlate to the part quality [4]. Melt pool width, for example, working closely with hatch distance, can predict void type defects between two adjacent tracks. Melt pool depth, on the other hand, determines the fusion between layers [5]. Figure 1(a) shows an ideal melting condition where the melted areas are well connected between tracks and layers. However, the un-melted powders would appear when the melt pool size doesn't reach the hatch distance and layer thickness. Figure 1(b) shows a lack-of-fusion defect created by insufficient overlap between two melt pool tracks, which leaves track-wise un-melted powders. Figure 1(c) shows the layer-wise un-melted powder between the current and previous layers caused by an insufficient melt pool depth. Either condition could increase porosity in the final part.

Figure 1 shows the defects due to small melt pool sizes. Nevertheless, oversized melt pools may lead to a different type of defects called over melting [6]. If a laser beam frequently re-melts an area with oversized melt pools, this area can be overheated and develops keyholing [7]. Keyhole is a defect that affects both current and future layers. Keyholing creates voids in the current layer, which can significantly affect the powder spreading of the next layer [8,9].

Though melt pool size is not a property of AM parts, it is a critical process measurement with salient features associated with part quality. Many process parameters can directly or indirectly affect melt pool size. For example, instant energy density has been proved being a major factor in manipulating melt pool size [10,11]. Melt pool formation is highly sensitive to energy density-related variables such as laser power, scan speed,

[*] Electronic address: yan.lu@nist.gov; Corresponding author

Yang, Zhuo; Lu, Yan; Yeung, Ho; Krishnamurty, Sundar. "3D Build Melt Pool Predictive Modeling for Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

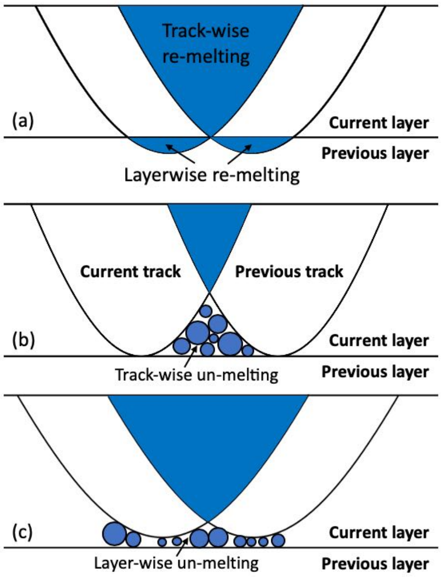


FIGURE 1: The cross-sectional view of two parallel melt pool tracks under three melting conditions. The solid blue area represents the re-melted area. The circles are the un-melted powder particles.

and hatch distance [12]. Energy density determines the maximum thermal input to the powder. Material properties such as heat absorption efficiency determine the actual thermal energy received by the powder [13]. Scan pattern, defining how a laser beam travels, is another important element that affects melt pool size. Experimental studies indicate that part built with the same process parameters but different scan patterns can have distinct thermal fields [14, 15].

If the melt pool size can be predicted before a part is built, users can modify the part design and process setting timely. Physics-based simulation approaches such as finite element analysis (FEA) and computational fluid dynamics (CFD) methods are mature and popular in AM research to predict melt pool size. Pioneer modeling works also reveal that melt pool formation is sensitive to both process parameters and scan patterns [14, 16]. An advantage of the FEA and CFD modeling methods is that they can simulate the effects of any physical parameter [17, 18], for example, various scan patterns. However, the high computational cost is hindering the use of physics-based modeling approaches. A single layer FEA simulation at millimeter-scale may take hours to calculate the interactions between multiple scan tracks. CFD models may be more expensive. This issue becomes noticeable when solving large scale AM problems. The computational cost also prevents the use of these simulations in model-based optimization, because it requires iterative runs to approach the optimal solution.

The data-driven modeling approach becomes a substitute for the physics-based modeling methods to achieve fast predictions. In general, data-driven modeling methods create black-box surrogate models based on experimental data. The computational time may vary from method to method but generally faster than complex simulations [19]. The prediction from surrogate model can usually be made within seconds [11, 20]. However, for data-driven methods, it could be very challenging to characterize process settings and formulate them into a model. For example, scan pattern involves long time series, numerous moving vectors, and dynamic laser spot locations. It is challenging to represent these features using a few variables. And from the perspective of data availability, most AM machines don't export the exact scan path of a build to the users [21].

The authors' previous works developed a data-driven method called Neighboring-Effect Modeling (NBEM) method which was able to address the challenge and predict the melt pool size for a single-layer [14]. However, there were multiple obstacles to further improvement of the model accuracy by extending the single layer to the multi-layers approach. First, it requires multi-layers in-situ melt pool monitoring data, which was not available. Besides, the NBEM factors that designed to capture the features of the single-layer scan strategy cannot quantify the variation between layers.

This study aims to address the research obstacles by developing a novel Layer-wise NBEM (L-NBEM) approach. The L-NBEM method introduces five additional input variables to predict melt pool size using the scan and exposure setting data from the previous layer. An experiment was conducted to collect build commands and melt pool images for multiple layers of one part. This data was used for model construction and validation. The following section mainly introduces the L-NBEM approach, including the physical hypothesis, the modeling variables, and the machine learning method for model training. Section 3 lists the details about the experiment and data. Section 4 presents the results of the L-NBEM model compared to other models. Section 5 discusses the result and future research plan.

2 LAYER-WISE NEIGHBORING-EFFECT MODELING METHOD

This section introduces the background, hypothesis, and modeling variables and method for the L-NBEM approach. The original NBEM method focuses on predicting the melt pool size for a single layer from the process parameters and scan pattern. The fundamental idea is to characterize the neighboring area process settings into two simplified NBEM variables. The proposed L-NBEM method builds on the previous method by introducing more variables to characterize the layer-wise features of process settings for multi-layers AM build.

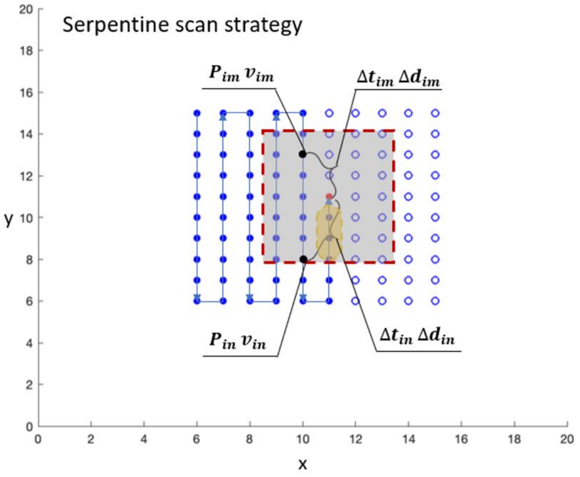


FIGURE 2: Neighboring points in serpentine scan strategy. [14]

2.1 Neighboring-Effect Modeling Method

The fundamental idea of the NBEM method is to use a minimum number of variables to characterize complex scan patterns and the corresponding process parameters. It assumes the melt pool at current laser spot is affected by the nearby region with a neighboring effect, which is a function of the scan pattern. Different scan patterns would visit the neighboring area of a laser scan spot differently. Thus, the time differences and the spatial differences between the previous spots and the current laser spot can be used to characterize the scan pattern.

Figure 2 shows an example of constructing an Ω matrix [14] to capture all the factors affecting the melt pool size within a neighbor region. The red, solid blue, and unfilled blue points in the figure represent the current, previous, and future scan points, respectively. The arrow marks the scan direction of the serpentine scan strategy. The gray area is the neighboring-effect zone enclosed by the red dashed box. This study investigates the neighboring area as a $0.1mm \times 0.1mm$ square. The solid blue points in the gray area are the neighboring points to be included in the Ω matrix. In this work, the points located on the same straight track of the current focal spot are not included in the matrix (dashed yellow area) to avoid over-estimation. It assumes that the laser power and scan speed at the current scan point can cover the effect from these points since it is generally a single track problem.

Parameters P_{im}, v_{im}, P_{in}, v_{in} are the laser power and scan velocity associate to point m and point n, respectively. Δt_{im} and Δd_{im} are the time and distance difference between former point m and current point i. Similarly, Δt_{in} and Δd_{in} are assigned to point n. In this example, $\Delta t_{im} > \Delta t_{in}$ and $\Delta d_{im} > \Delta d_{in}$. Subscript i denotes the i_{th} spot, which uses the global indexing. m and n are the local indexing referring to the i_{th} point. Collecting the variables for all neighboring points facilitates the formulation of the Ω_i for the current focal point i. Collecting the variables for all neighboring points except the points located in the yellow area facilitates the formulation of the Ω_i for the current focal point i. The dashed yellow area is considered as a single-track problem which is already covered by the laser power and scan speed of current point.

With Ω_i formulated, melt pool area can be expressed as a function of processing parameters and processing history. The smallest Ω_i is an empty matrix which is assigned to the first focal point on a toolpath since no previous points exist at that moment. Points in inner build areas usually have larger Ω_i than those located on the edge. The NBEM factors can be calculated by:

$$f(\Omega_i) = \begin{bmatrix} P_{i1}/v_{i1} \\ P_{i2}/v_{i2} \\ \vdots \\ P_{ij}/v_{ij} \end{bmatrix}^T \begin{bmatrix} f_{\Delta t}(\Delta t_{i1}) & f_{\Delta d}(\Delta d_{i1}) \\ f_{\Delta t}(\Delta t_{i2}) & f_{\Delta d}(\Delta d_{i2}) \\ \vdots & \\ f_{\Delta t}(\Delta t_{ij}) & f_{\Delta d}(\Delta d_{ij}) \end{bmatrix} \quad (1)$$

$\theta_i^{\Delta t}$ and $\theta_i^{\Delta d}$ represent the integrated factor of $f(\Omega_i)$ on time and distance perspective. $f_{\Delta t}(\Delta t_{ij}$ and $f_{\Delta d}(\Delta d_{ij}$ represent the scaling functions of time-lapse and distance, respectively. In Equation (2) and Equation (3), the input laser power P_{ij} of the neighboring point j provides a fundamental impact. Function $f_{\Delta t}(\Delta t_{ij})$ and $f_{\Delta d}(\Delta d_{ij})$ are used to scale the neighboring-effect from 0 (no impact) to 1 (strongest impact). Points geometrically and/or temporally remote to the current focal point are scaled to have minimal impact. Δt_{ij} or Δd_{ij} can be too large to provide any impact since an irradiated area is limited by its size.

$$\theta_i^{\Delta t} = \sum_{j=1}^{j=n} f_{\Delta t}(\Delta t_{ij} \frac{P_{ij}}{v_{ij}} \quad (2)$$

$$\theta_i^{\Delta d} = \sum_{j=1}^{j=n} f_{\Delta d}(\Delta d_{ij} \frac{P_{ij}}{v_{ij}} \quad (3)$$

The time-neighboring-effect focuses on modeling the preheating conditions of the current focal point, which depends on powder cooling rate. The cooling rate can affect the preheating temperature of melting. The literature indicates the temperature of the focal point decreases quickly at the beginning but the total time for cooling and irradiated area can be varied [22,23]. In this work, the time-neighboring-effect is formulated exponentially:

$$f_t(\Delta t_{ij}) = a_1 e^{a_2 \Delta t_{ij}} \quad (4)$$

where $f_t(0) = 1$ and $f_t(\Delta t_{max}) = 0$. The parameters a_1, a_2 and Δt_{max} is a fixed value which is derived from experiments.

The distance-neighboring-effect aims to formulate the spatial impact such as spattering and denuded powder. Simulation and experimental results indicate the denuded width of an irradiated area is ranged from 0.04 mm to 0.06 mm for high-resolution scanning [24,25]. The distance-neighboring-effect is formulated as:

$$f_d(\Delta d_{ij}) = b_1 e^{b_2 \Delta d_{ij}} \quad (5)$$

where $f_d(0) = 1$ and $f_d(\Delta d_{max}) = 0$. Optimal coefficient b_1 and b_2 will be determined by experimental data. Δt_{max} and Δd_{max} equal to 20 ms and 0.6 mm in this study. The melt pool area, $\tilde{y}_i$, is a function of laser power at current spot, P_i, current scan speed, v_i, and NBEM time and spatial factors $\theta_i^{\Delta t}$ and $\theta_i^{\Delta d}$.

$$\tilde{y}_i = f(P_i, v_i, \theta_i^{\Delta t}, \theta_i^{\Delta d}) \quad (6)$$

2.2 Hypothesis of L-NBEM

In general, melt pool size depends on the energy absorbed by the powder. Higher energy density imported to the powder can make the metal powder particles melted to liquid and fuse to the nearby area easier [26, 27]. Once the metal liquid contacts the powder outside the laser spot, if there is enough additional energy, it would melt more metal powder thus enlarge the melt pool. According to this phenomenon, melt pool size typically is an outcome of thermal energy input. The following hypothesis is made based on this finding.

This study generated the data using the same testbed at one AM build with virgin metal powder. It assumes the machinery and environmental conditions remain the same during the entire process. The model would ignore the variance of powder particle size, changing of chamber temperature and humidity, and fluctuation of laser power. The L-NBEM method of this study considers them as constant parameters for all layers.

Simulations and experiments indicate that the melt pool size is a product of energy input. Laser power and scan speed are two significant components of energy density [28, 29]. The NBEM method includes these two major variables according to the importance of them relates to the energy input. Given the same energy input, however, the melt pool size can be changed due to different preheating temperatures [30]. Higher initial temperature establishes a preheating condition of the powders thus generate larger melt pool size. As a result, factors that may affect the preheating temperature should be included. The track-wise factors in the same layer are characterized by the NBEM method. L-NBEM mainly introduces additional layer-wise factors. The total energy input and cooling time on the previous layer determine the preheating temperature of the current layer. Generally, it is a heat accumulation and releasing process.

The first layer of the part usually builds on the bare build plate with relatively ideal conditions of surface roughness and uniform chamber temperature. The single-track experiment shows the melt pool under the same process parameters produce fewer uncertainties on the bare plate than coarse powders [29]. However, the powders for later layers are spread on the previous layer unless overhanging occurs. Therefore, the L-NBEM method assumes the melting conditions of the previous layer can affect current melt pool formation. To cover this hypothesis, the L-NBEM method would characterize the features of a specific field on the previous layer. This field locates at the projection area of the current NBEM area. Those features from the layer-wise affect to the melt pool size.

Another hypothesis of L-NBEM method is the current layer can only be impacted by the most recent layer. It assumes the layer-wise melt pool features has already covered the thermal history of all former layers.

2.3 Overview of L-NBEM

The L-NBEM method divides the potential factors of the melt pool size into two groups. The first group includes the track-wise factors on the current layer within the NBEM region, the red dashed area in Figure 3. The second group includes the layer-wise factors on the previous layer within the projection of the NBEM region. Figure 3 shows an example of NBEM and L-NBEM regions on Layer 9 and Layer 8. Both layers using the serpentine scan pattern. Arrow lines show the scan path from the beginning to the end. The red and black arrows represent scanned tracks and future tracks, respectively. The orange square represents the NBEM region. Historical points within the NBEM region would be used to calculate the track-wise factors. The red dot located in the center is the current laser spot. The surface plot of Layer 8 represents the melt pool area map. The colormap ranges from 0 mm^2 to 0.04 mm^2. The orange box is the projection area of NBEM region, which crops the L-NBEM region. Historical points located within the L-NBEM region would be used to calculate the layer-wise variables.

Five NBEM factors are formulated to the track-wise input variables: building time from start point to the laser spot (t_i), laser power (P_i), scan speed (v_i) at the laser spot, NBEM time factor ($\theta_i^{\Delta d}$), and NBEM distance factor ($\theta_i^{\Delta d}$). The layer-wise input variables that represent the effect from the previous layer

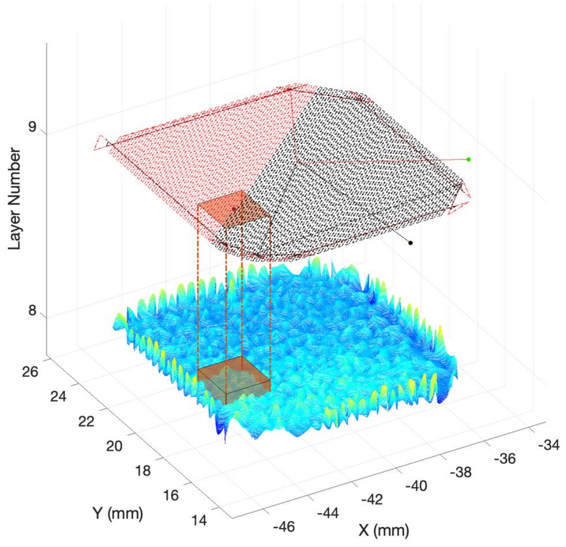


FIGURE 3: An example of two neighboring layers using serpentine scan pattern.

are: total energy input on the previous layer (J), laser idle time from the end of the previous layer to start of current layer (λ), mean (A_{avg}), maximum (A_{max}), and standard deviation (A_{var}) of the melt pool area of the L-NBEM region. The general formulation of the L-NBEM model to predict the melt pool area $\tilde{y}_i$ at current laser spot can be presented as:

$$\tilde{y}_i = f(t_i, P_i, v_i, \theta_i^{\Delta t}, \theta_i^{\Delta d}, J, \lambda, A_{avg}, A_{max}, A_{var}) \tag{7}$$

Figure 4 is the 2D view of Figure 3 when stacking the layers into a 2D plot. If the laser spot located on edge, NBEM and L-NBEM would have fewer points. This study uses a square box to filter NBME and L-NBEM fields. However, there is no limitation on the shape of the region.

2.4 Input and Output of L-NBEM

Table 1 lists the input variables in Equation (7) for their unit and function. Variable with star sign indicates it represents the layer-wise effect. Pound sign marks the dependent variable. Energy input is the total energy input from the previous layer, which is a production of laser power at each time step and the total time step. This study sets a constant time step to 10 μs. Thus, the total energy input of any layer could be presented as:

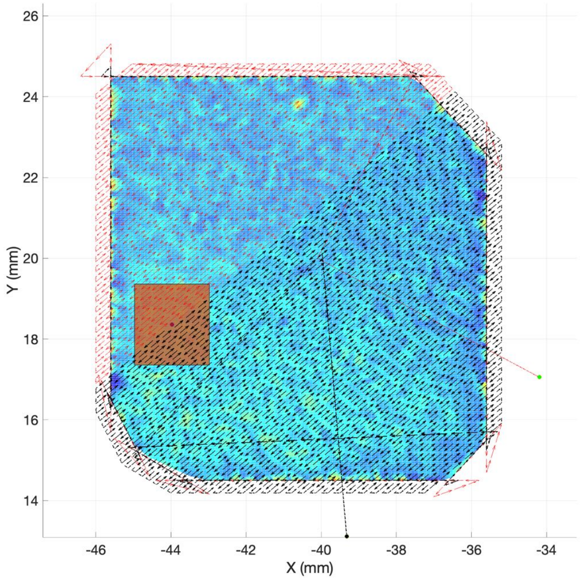


FIGURE 4: 2D view shows the top view of Figure 3 that visualize the layer-wise and track-wise effect when stack two layers together.

$$J = \sum_{i=1}^{n} P_i \tag{8}$$

Where P_i is the laser power for the i_{th} laser spot, n is the total laser spots for one layer. In fact, the laser spots are not discrete points since they are physically continuously connected. This work uses the time interval ($10 \mu s$) from the digital commands to separate the laser spots. Thus, the index i of the laser spot is the same to the time step.

Figure 5 shows three examples of melt pool coaxial images. These images were taken at different locations of one layer with same laser power and scan speed. Figure 5(a) is the melt pool less than regular size. Figure 5(b) is the melt pool with average size. Figure 5(c) is the melt pool with very large area. This study chooses area as the output to represent the melt pool size. The grayscale threshold 100 is used to find the contour of the melt pool. The pixels within the contour are considered as the melt pool. The melt pool coaxial image is 120 × 120 pixels, where each pixel is a 8 $\mu m \times 8$ μm square.

Figure 6 shows the melt pool area measured by different threshold values. Figure 6(a) is the original melt pool coaxial image. Figure 6(b) shows the melt pool size by grayscale thresholding set at, 100, 120, and 180. The smallest melt pool area,

Yang, Zhuo; Lu, Yan; Yeung, Ho; Krishnamurty, Sundar. "3D Build Melt Pool Predictive Modeling for Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

TABLE 1: Name, symbol, unit, and function of the variables. The pound sign (#) marks the dependent variables. The star sign (*) marks the variables representing the layer-wise effect.

Variable name	Symbol	Unit	Function
Building time	t_i	ms	Calculate the cumulative heat
Laser power	P_i	W	Input heat source
Scan speed	v_i	mm/s	Affect the energy density at the laser spot
NBEM time #	$\theta_i^{\Delta t}$	N/A	Characterize the time effect
NBEM distance #	$\theta_i^{\Delta d}$	N/A	Characterize the spatial effect
Energy input *, #	J	W	Calculate the total energy input from previous layer
Idle time *	λ	ms	Calculate the cooling time
Mean area *, #	A_{avg}	mm^2	Calculate average melting conditions of previous layer
Maximum area *, #	A_{max}	mm^2	Calculate extreme melting conditions of previous layer
Standard deviation of area *, #	A_{var}	mm^2	Evaluate the variation of melt pool of previous layer

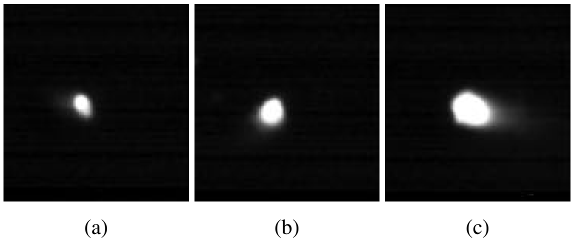


FIGURE 5: Sample melt pool coaxial images taken at different locations.

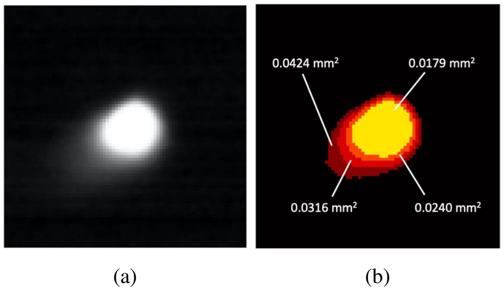


FIGURE 6: Melt pool area measurement based on different threshold grayscale value.

0.0179 mm^2, is derived based on the highest thresholding 180. The largest melt pool area, 0.0424 mm^2, is measured using the lowest thresholding 80. These two values represent the wrong melt pool size since both thresholding is out of the range found physically. This study selects 100 as the threshold value to measure all the melt pool images. The selected threshold may not be the most accurate number to find the actual melt pool. However, it can represent the melt pool size changes caused by scan pattern and process parameters if all the measurements using the same criteria. Specific to the example shown in Figure 6, the area is 0.0316 mm^2 by grayscale threshold 100.

2.5 Modeling

A feedforward neural network (NN) model is trained to represent the L-NBEM model. Figure 7 plots the structure of the neural network that include input layer, hidden layer, and output layer. The 10 variables in Equation 7 construct the input layer. The hidden layers contain two fully connected layers with 20 nodes for each. Melt pool area constructs the output layer. Levenberg-Marquardt is the activation function. For comparison purpose, polynomial regression is used to build the model by traditional power-speed method.

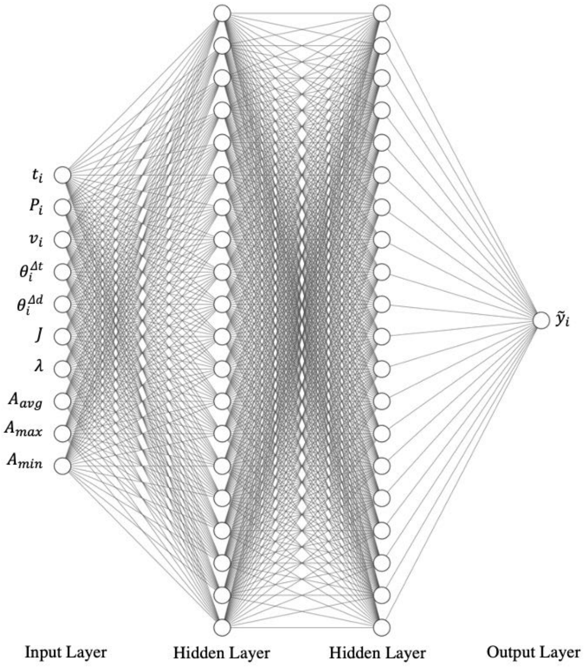


FIGURE 7: Structure of the Neural Network.

2.6 Validation

This study uses the leave-one-out cross-validation (LOOCV) method to validate the proposed method. The data of the layer being validated would not be included in the training dataset for each layer. It would be the validation dataset to validate the model for this layer. An n-layers problem would establish n models for LOOCV.

The criteria for performance evaluation are Average Relative Error Magnitude (AREM) and Average maximum Error Magnitude (Average-MREM) [20]. AREM can represent the average error for all predictions of one particular layer.

$$AREM = \frac{\sum_{i=1}^{m} |y_i - \tilde{y}_i|}{my_i} \qquad (y_i \neq 0) \qquad (9)$$

where m is the total number of validation data points. Parameters y_i and $\tilde{y}_i$ are the actual observation and prediction value of the melt pool area. Since only positive laser power can create melt pool and this experiment always turn on the laser during the build, this study would not have divisor equal to zero.

The average-MREM calculates the average error of the 100 largest error points of each layer. This method aims to evaluate the performance of L-NBEM for extreme conditions. The MREM formula for one layer is:

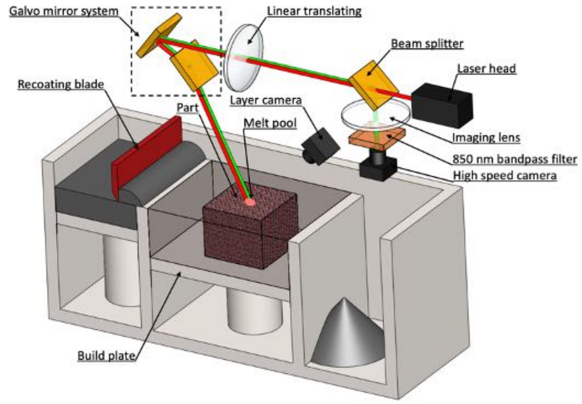


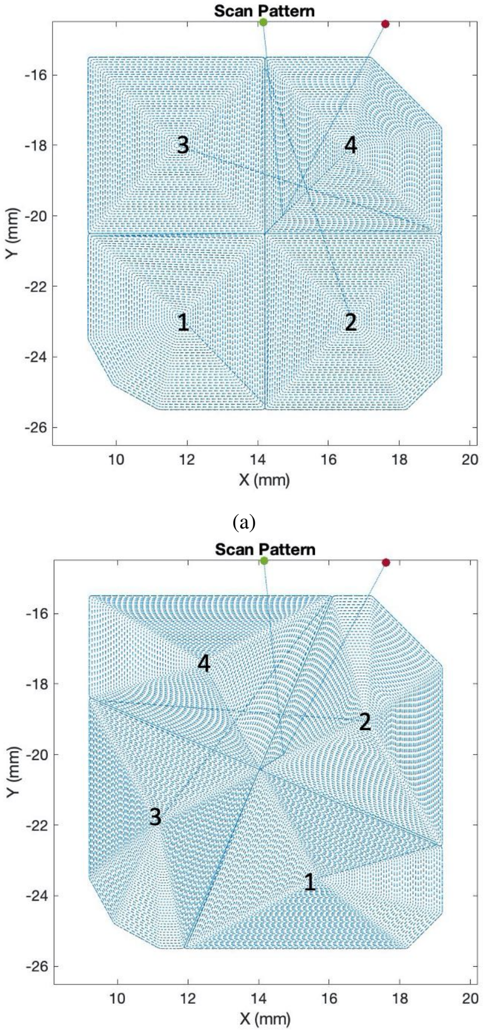
FIGURE 8: The conceptual model of AMMT [14].

$$MREM = max(\frac{|y_i - \tilde{y}_i|}{y_i}) \qquad (y_i \neq 0) \qquad (10)$$

3 EXPERIMENT DESIGN

The experiment is conducted on the Additive Manufacturing Metrology Testbed (AMMT) at National Institute of Standards and Technology (NIST) as shown in Figure 8. The AMMT [25] is a fully customized metrology instrument that enables flexible control and measurement of the Laser Powder Bed Fusion process. An in-house developed AM software (SAM), which is capable of stereolithography (STL) slicing, scan path planning, G code generation and interpretation [31], was used to program the different scan strategies for the experiment. Inconel 625 powder and substrate were used, where the substrate has a dimension of 101.6 $mm \times$ 101.6 $mm \times$ 12.7 mm. Twelve rectangular parts (with chambered corners) of dimensions 10 $mm \times$ 10 $mm \times$ 5 mm were laid on the substrate, with a minimum spacing of 10 mm between the parts. Each part was built with a different scan strategy. The melt pool was monitored by a high-speed camera which is optically aligned with the heating laser, such that the image of the melt pool is maintained stationary within the camera's field of view. The camera was triggered at every 200 μs (i.e., 2000 frames per second), with an integration time of 20 μs.

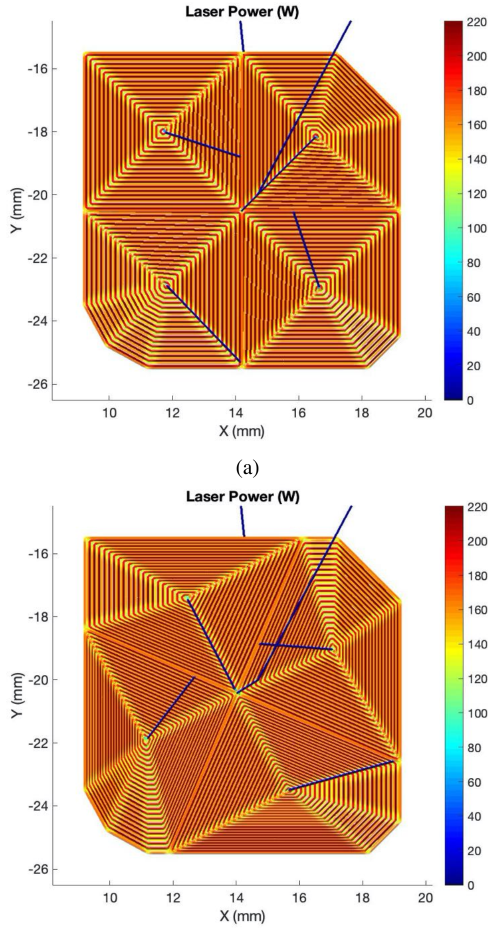
The experiment applies the 'island' spiral concentrating scan strategy to build the part. The part has 250 layers where each layer is 20 μm. To avoid high heat concentration and introduce variance of the island shape, the machine would rotate the centroid angle at each layer. The rotation angle between layers is 83.4 degree and the first layer divide the islands in the vertical intersection as shown in Figure 9(a). After rotating the intersection for 83.4 degrees, the scan pattern for Layer 2 is shown in



(a)

(b)

FIGURE 9: Scan pattern for Layer 1 (a) and Layer 2 (b) of the part. The scan starts at green point and finishes at the red point. Numbers marks the scan order for each island



(a)

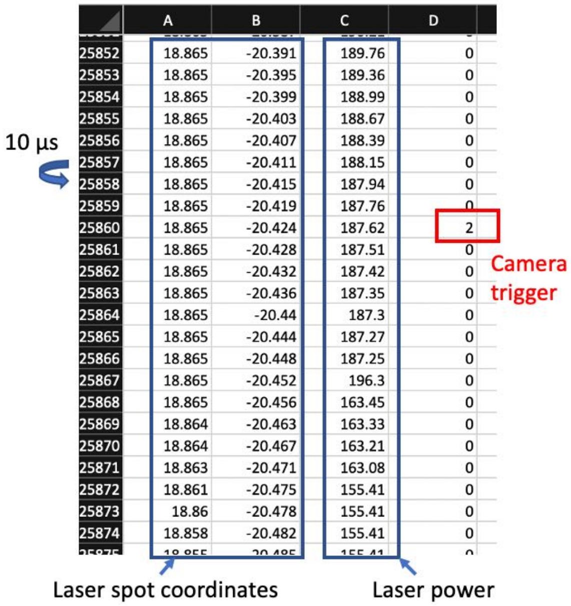
(b)

FIGURE 10: Laser power for Layer 1 and Layer 2

Figure 9(b). For each layer, the laser beam starts at the green point and firstly scan the contour of the part and each island. The laser then moves from the edge spirally concentrating to the center of the first island. After finishing the first island at the island center, it moves to the edge of the second island.

Figure 10 shows the laser power of Layer 1 and Layer 2. The dark blue lines represent when laser traveling between islands with no power input. The machine reduces the scan speed when laser turning direction. The machine reduces the layer power for scan speed to avoid high energy input. The range of laser power and scan speed is from 0W to 234.83 W and 0 to 900 mm/s, respectively.

The AMMT uses XY2-100 control system to control the laser scan [31]. It specifies the laser process parameters (laser coordinates, power, and camera trigger) at each time step. Figure 11 shows the digital commands in a spreadsheet from step 25852 to 25874. Column A and Column B have the laser spot position on the x-axis and y-axis. Column C is the laser power. Column D identifies the coaxial camera trigger. The camera would capture an in-situ image when the value change to 2. The time interval between two steps is 10 μs. The scan speed is calculated from the position between two steps.



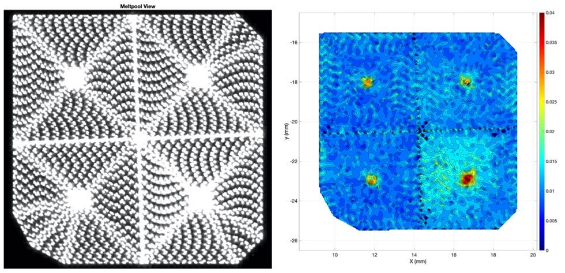
	A	B	C	D
25852	18.865	-20.391	189.76	0
25853	18.865	-20.395	189.36	0
25854	18.865	-20.399	188.99	0
25855	18.865	-20.403	188.67	0
25856	18.865	-20.407	188.39	0
25857	18.865	-20.411	188.15	0
25858	18.865	-20.415	187.94	0
25859	18.865	-20.419	187.76	0
25860	18.865	-20.424	187.62	2
25861	18.865	-20.428	187.51	0
25862	18.865	-20.432	187.42	0
25863	18.865	-20.436	187.35	0
25864	18.865	-20.44	187.3	0
25865	18.865	-20.444	187.27	0
25866	18.865	-20.448	187.25	0
25867	18.865	-20.452	196.3	0
25868	18.865	-20.456	163.45	0
25869	18.864	-20.463	163.33	0
25870	18.864	-20.467	163.21	0
25871	18.863	-20.471	163.08	0
25872	18.861	-20.475	155.41	0
25873	18.86	-20.478	155.41	0
25874	18.858	-20.482	155.41	0
25875	18.855	-20.485	155.41	0

FIGURE 11: The XY2-100 digital commands from step 25852 to step 25874. The column on the left is the time step.

The experiment builds 12 parts at the same time that placed on different positions on the build plate. In other words, the machine needs to scan 12 parts at different locations of each layer. The camera, however, can only focus on one part of each layer. As a result, the experiment collects the melt pool data for the particular part every 12 layers. Finally, a total of 21 layers of data are available for training and validating the L-NBEM model. It include 6,192,495 rows of digital command data for 21 layers and 118,928 melt pool in-situ images. The melt pool area is calculated from these in-situ images using grayscale threshold value 100.

4 RESULT

This section compares the L-NBEM prediction performance with the traditional power-speed model and the NBEM model. For visualization, the melt pool area is mapped into contour plots. All the contour plots, for both measurement and prediction, use the same colormap that ranged from 0 (dark blue) to 0.04 (red) mm^2. Figure 12 shows an example of the transmission from the melt pool images to the melt pool area map. Figure 12(a) distributes all 4,589 melt pool in-situ images of Layer 177 in one map at the position where the image taken. The melt pool area is calculated using a threshold of gray scale 100. Figure 12(b) uses the measured melt pool data to create the contour plot. The melt pool size changes at different locations. For example, the center of each island has larger area than island average. The bottom right island has largest melt pool out of the entire layer.



(a)

(b)

FIGURE 12: (a) is the map of melt pool in-situ images of Layer 177. (b) is the melt pool area contour plot of this layer.

Figure 13 shows the contour plot from Layer 201 to Layer 249. The solid red box on the left lists the scan pattern and process parameters of Layer 249. The dashed red box on the right lists the scan pattern of Layer 213. Layers present different features on melt pool size, which is mainly caused by the changes in scan strategy. The objective of the L-NBEM model is to predict the variance between layers caused by track-wise and layer-wise variables.

To visualize the result, Figure 14 stacks all the measured layers in one 3D view with a 0.5 transparent ratio. As shown in the figure, the melt pool area average, island average, and oversized melt pool field present significant differences between layers. Some layers have an average area of around 0.02 mm^2. However, others can be as low as 0.015 mm^2. The oversized melt pool is highly clustered at the island center. Furthermore, due to the rotation rule of the island division between layers, the red area twisted from the bottom layer to the top layer. Figure 14 shows the contour plot of the measured melt pool area for all available 21 layers. The colormap is ranged from 0 to 0.04 mm^2.

The melt pool measurement is first compared to the prediction by the traditional power-speed model using the polynomial regression method. The polynomial regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. The power-speed model is popular and reliable in solving most single track problems [29]. The formulation of the model is:

$$\tilde{y}_i = f(P_i, v_i) \tag{11}$$

Figure 15 shows the contour plot of the melt pool predicted from the power-speed model. The AREM and Average-MREM of the LOOCV is 19.41% and 77.30%. As shown in the figure,

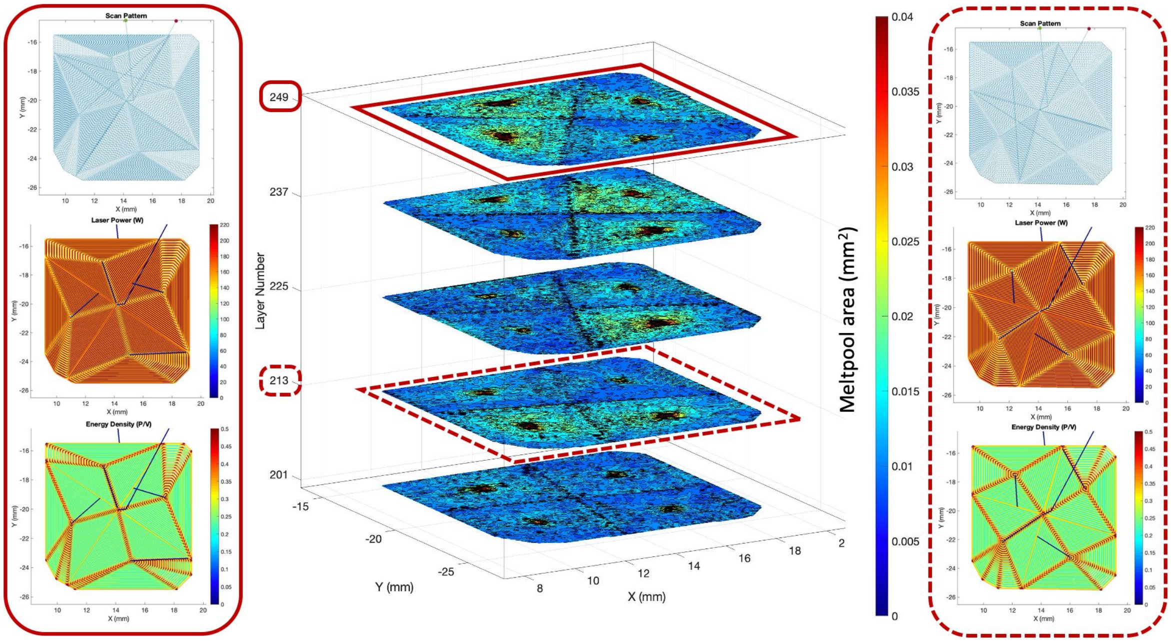


FIGURE 13: The contour plot from Layer 201 to 249. The solid and dashed boxes show the scan pattern for Layer 213 and 249.

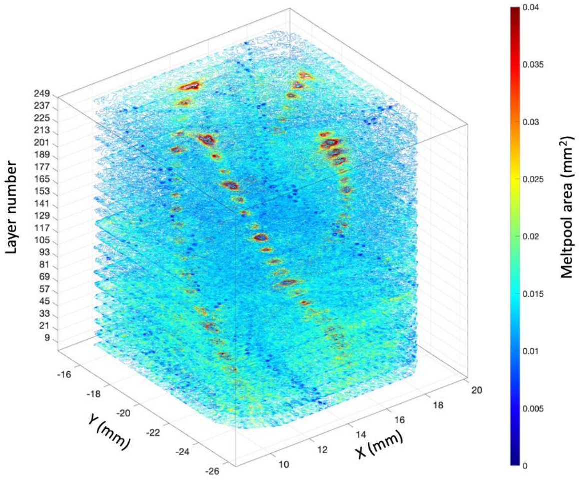


FIGURE 14: The contour plot of the measured melt pool area for all 21 layers.

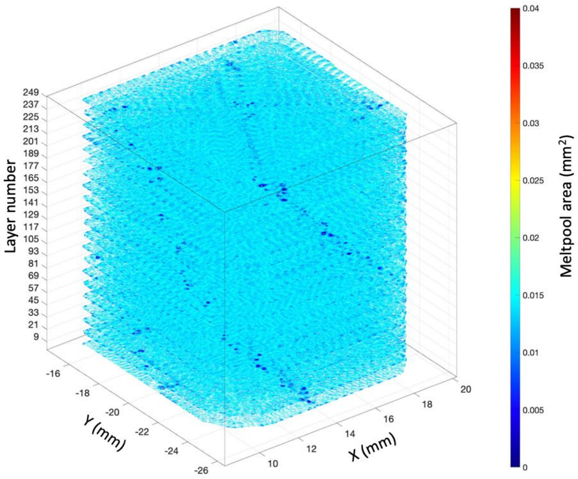


FIGURE 15: The contour plot of the melt pool prediction by power-speed model.

the prediction of each layer is very different to the actual melt pool area. It can predict some fields with relatively small melt pool due to insufficient energy input, the darker blue area. However, the major issue of this model is that it cannot distinguish the difference caused by the scan pattern, such as irregular melt pool field at the island center.

For comparison, the NBEM method is used to single layer model for all 21 layers. It is also the first time to test this approach using multi-layers experimental data. The NBEM model, which was designed for solving single layer problems, can pre-

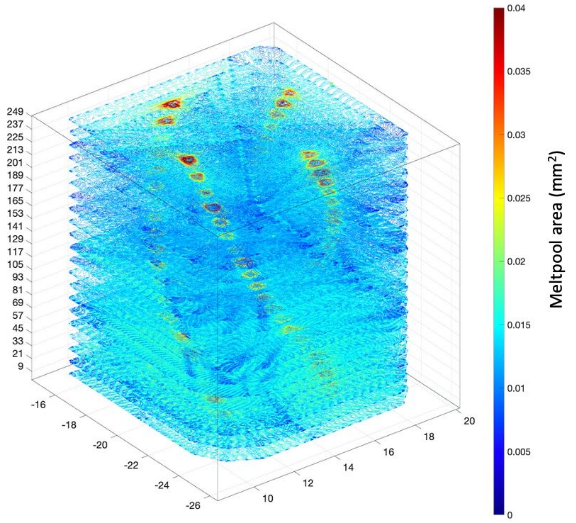


FIGURE 16: The contour plot of the melt pool size prediction by the NBEM model.

dict the irregular melt pool area due to scan pattern changes. The NBEM model is trained using the quadratic polynomial regression method. The AREM and Average-MREM of LOOCV is 15.23% and 70.3%, respectively.

Figure 16 shows the melt pool area predicted by the NBEM method. Compared to Figure 14, the model can predict the regular and irregular melt pool located on the layers. The differences between layers are also reflected on the contour plots. However, due to a lack of consideration for the layer-wise effect, predictions for the the details remain an issue.

Figure 17 shows the prediction resulted from the L-NBEM model. The LOOCV result of this method presents the lowest AREM and average-MREM, 12.12% and 64.13%, respectively. The contour plot based on the L-NBEM prediction is the closest to the actual measurement compared to the power-speed and NBEM models.

Figure 18 shows the LOOCV AREM for the melt pool predictions using the power-speed, NBEM, and L-NBEM models based on the data for all the 21 layers. L-NBEM has the lowest global average AREM. L-NBEM also demonstrates the lowest AREM for 19 layers. However, Layer 69 of L-NBEM presents the largest AREM (22.52 %) while the NBEM model presents the lowest AREM (14.22%) for Layer 213.

5 DISCUSSION AND FUTURE WORK

The objective of this work is to extend the NBEM method from a single layer approach to a multi-layer approach for melt pool size prediction. For this purpose, this study introduces five

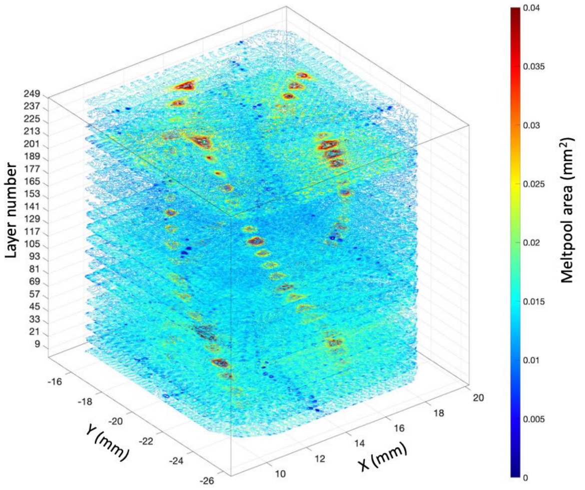


FIGURE 17: The contour plot of the melt pool prediction by the L-NBEM model.

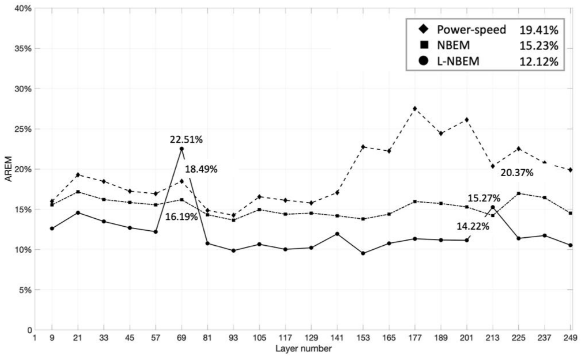


FIGURE 18: LOOCV AREM for all 21 layers of power-speed (P-v), NBEM, and L-NBEM methods.

additional variables to enhance the model performance. Generally speaking, the track-wise variables captured in the NBEM model characterize the effect from the scan pattern of the current layer. The newly added layer-wise variables characterize the effect of the previous layers. By introducing these new variables, L-NBEM can predict the melt pool size better for 3D AM builds.

An experiment provides more than 100,000 coaxial melt pool in-situ images and 6 million high sampling digital commands to validate the effectiveness of the proposed approach. All the data are deployed for training and validating the L-NBEM model using the LOOCV method. The L-NBEM model shows the lowest AREM compared to both the traditional power-speed and the NBEM models. The L-NBEM model can predict both regular melt pools that following general energy density rules and irregular melt pools introduced by the scan pattern. The model may be helpful for layerwise feedback process control of LPBF machines.

However, it is also observed that the L-NBEM model cannot guarantee a predictive accuracy for all layers. For example, Layer 69 is an outstanding layer that shows large predictive errors. The AREM of that layer reaches 22.51% which is higher than that of the NBEM model and that of the power-speed model. The error could be introduced because of the neural network construction. Various neural network configurations were tried, such as changing the number of hidden layers and neurons, different functions, and different parameters. However, none of them improved the results significantly. Layers with outstanding prediction errors exist and show up randomly. In the future, a more comprehensive machine learning approach would be investigated to reduce the modeling inconsistency. Model uncertainty quantification should be considered since such modeling uncertainties may significantly affect the melt pool size prediction [32, 33].

The prediction error may also be generated from the experimental method. The part was built with 250 thin layers. However, due to a data acquisition limitation, the in-situ melt pool monitoring data was captured every 12 layers. Because of that, some layer-wise variables were calculated from estimates instead of direct measurements. This could lead to many data errors prior to model training. A more precise experimental design may help address this issue.

Future work would focus on building the correlation between melt pool size and final material properties such as porosity and residual stress. When preparing this paper, the co-authors were working on collecting the ex-situ data of the parts using X-ray computed tomography (XCT) scan. The preliminary findings based on the XCT scan data indicate that the distribution of the voids is similar to the melt pool area distribution. As shown in Figure 19, larger melt pool size regions (measured) usually locate at the center of each island at every layer. Due to the island division rotation strategy, these regions form a circle after projecting all the layers into a 2D figure. Coincidentally, the preliminary result of the XCT data shows the voids of the part in a similar pattern. A physics based explanation is that large melt pool size regions reflect the overheating at the island center during the process. The overheating may create keyholes and finally produce the voids. Future work would investigate the potential correlation between melt pool size and porosity. If part porosity and melt pool size have a great correlation, it would be possible to predict voids using in-situ melt pool images.

6 DISCLAIMER

Certain commercial systems are identified in this paper. Such identification does not imply recommendation or endorsement by NIST; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

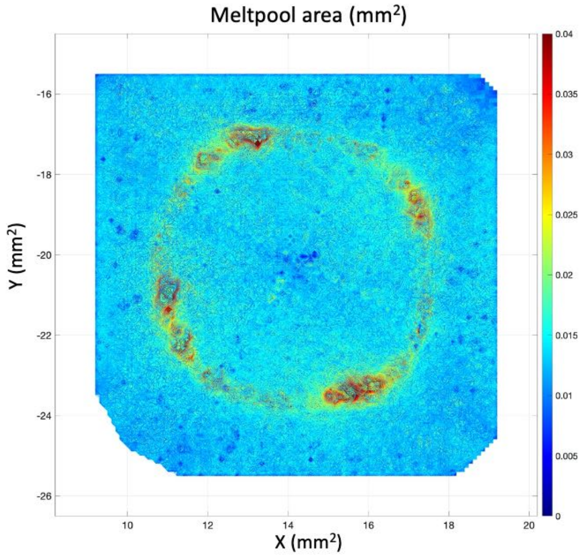


FIGURE 19: 2D top view for the measure melt pool contour plot after stacking all 21 layers together.

ACKNOWLEDGMENT

This material is based upon work supported by the National Institute of Standards and Technology (NIST) under Cooperative Agreement number NIST 70NANB18H258.

REFERENCES

[1] King, W., Anderson, A. T., Ferencz, R. M., Hodge, N. E., Kamath, C., and Khairallah, S. A., 2015. "Overview of modelling and simulation of metal powder bed fusion process at lawrence livermore national laboratory". *Materials Science and Technology,* ***31***(8), pp. 957–968.

[2] Luo, N., Scheitler, C., Ciftci, N., Galgon, F., Fu, Z., Uhlenwinkel, V., Schmidt, M., and Körner, C., 2020. "Preparation of fe-co-b-si-nb bulk metallic glasses by laser powder bed fusion: Microstructure and properties". *Materials Characterization,* ***162***, p. 110206.

[3] Matilainen, V., Piili, H., Salminen, A., Syvänen, T., and Nyrhilä, O., 2014. "Characterization of process efficiency improvement in laser additive manufacturing". *Physics Procedia,* ***56***, pp. 317–326.

[4] Frazier, W. E., 2014. "Metal additive manufacturing: a review". *Journal of Materials Engineering and Performance,* ***23***(6), pp. 1917–1928.

[5] Gong, H., Rafi, K., Gu, H., Starr, T., and Stucker, B., 2014. "Analysis of defect generation in ti–6al–4v parts made using powder bed fusion additive manufacturing processes". *Additive Manufacturing,* ***1***, pp. 87–98.

[6] Fox, J. C., Lopez, F., Lane, B. M., Yeung, H., and Grantham, S., 2016. "On the requirements for model-based thermal control of melt pool geometry in laser powder bed fusion additive manufacturing". In Proceedings of the 2016 Material Science & Technology Conference, Salt Lake City, pp. 133–140.

[7] King, W. E., Barth, H. D., Castillo, V. M., Gallegos, G. F., Gibbs, J. W., Hahn, D. E., Kamath, C., and Rubenchik, A. M., 2014. "Observation of keyhole-mode laser melting in laser powder-bed fusion additive manufacturing". *Journal of Materials Processing Technology,* ***214***(12), pp. 2915–2925.

[8] Matilainen, V.-P., Piili, H., Salminen, A., and Nyrhilä, O., 2015. "Preliminary investigation of keyhole phenomena during single layer fabrication in laser additive manufacturing of stainless steel". *Physics Procedia,* ***78***, pp. 377–387.

[9] Khairallah, S. A., Anderson, A. T., Rubenchik, A., and King, W. E., 2016. "Laser powder-bed fusion additive manufacturing: Physics of complex melt flow and formation mechanisms of pores, spatter, and denudation zones". *Acta Materialia,* ***108***, pp. 36–45.

[10] Lopez, F., Witherell, P., and Lane, B., 2016. "Identifying uncertainty in laser powder bed fusion additive manufacturing models". *Journal of Mechanical Design,* ***138***(11).

[11] Yang, Z., Hagedorn, T., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., Lu, Y., and Witherell, P., 2017. "A domain-driven approach to metamodeling in additive manufacturing". In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[12] Ciurana, J., Hernandez, L., and Delgado, J., 2013. "Energy density analysis on single tracks formed by selective laser melting with cocrmo powder material". *The International Journal of Advanced Manufacturing Technology,* ***68***(5-8), pp. 1103–1110.

[13] Devesse, W., De Baere, D., and Guillaume, P., 2014. "The isotherm migration method in spherical coordinates with a moving heat source". *International Journal of Heat and Mass Transfer,* ***75***, pp. 726–735.

[14] Yang, Z., Lu, Y., Yeung, H., and Krishnamurty, S., 2020. "From scan strategy to melt pool prediction: A neighboring-effect modeling method". *Journal of Computing and Information Science in Engineering,* ***20***(5).

[15] Gibson, I., Rosen, D. W., Stucker, B., et al., 2014. *Additive manufacturing technologies*, Vol. 17. Springer.

[16] Beal, V., Erasenthiran, P., Hopkinson, N., Dickens, P., and Ahrens, C., 2006. "The effect of scanning strategy on laser fusion of functionally graded h13/cu materials". *The International Journal of Advanced Manufacturing Technology,* ***30***(9-10), pp. 844–852.

[17] Yan, W., Ge, W., Smith, J., Wagner, G., Lin, F., and Liu, W. K., 2015. "Towards high-quality selective beam melting technologies: modeling and experiments of single track formations". In 26th Annual international symposium on solid freeform fabrication, Austin, Texas.

[18] King, W. E., Anderson, A. T., Ferencz, R. M., Hodge, N. E., Kamath, C., Khairallah, S. A., and Rubenchik, A. M., 2015. "Laser powder bed fusion additive manufacturing of metals; physics, computational, and materials challenges". *Applied Physics Reviews,* ***2***(4), p. 041304.

[19] Shao, T., 2007. *Toward a structured approach to simulation-based engineering design under uncertainty*. University of Massachusetts Amherst.

[20] Yang, Z., Eddy, D., Krishnamurty, S., Grosse, I., Denno, P., and Lopez, F., 2016. "Investigating predictive metamodeling for additive manufacturing". In ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[21] Yeung, H., Neira, J., Lane, B., Fox, J., and Lopez, F., 2016. "Laser path planning and power control strategies for powder bed fusion systems". In The Solid Freeform Fabrication Symposium, pp. 113–127.

[22] Manvatkar, V., De, A., and DebRoy, T., 2014. "Heat transfer and material flow during laser assisted multi-layer additive manufacturing". *Journal of Applied Physics,* ***116***(12), p. 124905.

[23] Bertoli, U. S., Guss, G., Wu, S., Matthews, M. J., and Schoenung, J. M., 2017. "In-situ characterization of laser-powder interaction and cooling rates through high-speed imaging of powder bed fusion additive manufacturing". *Materials & Design,* ***135***, pp. 385–396.

[24] Hooper, P. A., 2018. "Melt pool temperature and cooling rates in laser powder bed fusion". *Additive Manufacturing,* ***22***, pp. 548–559.

[25] Lane, B., Mekhontsev, S., Grantham, S., Vlasea, M., Whiting, J., Yeung, H., Fox, J., Zarobila, C., Neira, J., McGlauflin, M., et al., 2016. "Design, developments, and results from the nist additive manufacturing metrology testbed (ammt)". In Solid Freeform Fabrication Symposium, Austin, TX, pp. 1145–1160.

[26] Ly, S., Rubenchik, A. M., Khairallah, S. A., Guss, G., and Matthews, M. J., 2017. "Metal vapor micro-jet controls material redistribution in laser powder bed fusion additive manufacturing". *Scientific reports,* ***7***(1), pp. 1–12.

Yang, Zhuo; Lu, Yan; Yeung, Ho; Krishnamurty, Sundar. "3D Build Melt Pool Predictive Modeling for Powder Bed Fusion Additive Manufacturing." Paper presented at ASME 2020 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference (IDETC/CIE 2020), St. Louis, MO, US. August 16, 2020 - August 19, 2020.

[27] Lane, B., Moylan, S., Whitenton, E. P., and Ma, L., 2016. “Thermographic measurements of the commercial laser powder bed fusion process at nist”. *Rapid prototyping journal*.

[28] Foroozmehr, A., Badrossamay, M., Foroozmehr, E., and Golabi, S., 2016. “Finite element simulation of selective laser melting process considering optical penetration depth of laser in powder bed”. *Materials & Design,* ***89***, pp. 255–263.

[29] Lu, Y., Yang, Z., Eddy, D., and Krishnamurty, S., 2018. “Self-improving additive manufacturing knowledge management”. In ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers Digital Collection.

[30] Vasinonta, A., Beuth, J. L., and Griffith, M., 2007. “Process maps for predicting residual stress and melt pool size in the laser-based fabrication of thin-walled structures”.

[31] Yeung, H., Lane, B. M., Donmez, M., Fox, J. C., and Neira, J., 2018. “Implementation of advanced laser control strategies for powder bed fusion systems”. *Procedia Manufacturing,* ***26***, pp. 871–879.

[32] Moges, T., Ameta, G., and Witherell, P., 2019. “A review of model inaccuracy and parameter uncertainty in laser powder bed fusion models and simulations”. *Journal of manufacturing science and engineering,* ***141***(4).

[33] Moges, T., Yan, W., Lin, S., Ameta, G., Fox, J., and Witherell, P., 2018. “Quantifying uncertainty in laser powder bed fusion additive manufacturing models and simulations”. In Solid Freeform Fabrication Symposium An Additive Manufacturing Conference.

Using Text Analytics Solutions with Small to Medium Sized Manufacturers: Lessons Learned

Michael Brundage[1*] and Radu Pavel[2]
[1] National Institute of Standards and Technology Gaithersburg, MD, USA
[2] TechSolve Inc. Cincinnati, OH, USA
Michael.brundage@nist.gov, pavel@TechSolve.org

Abstract

As Smart Manufacturing becomes more prevalent throughout industry, manufacturers are continuing to look for ways to more efficiently apply advanced data analysis methods to improve their decision processes. One promising area for improving decision making is through the use of natural language processing (NLP) methods on text-based data in maintenance. Maintenance personnel often capture important information on the problems and repairs throughout the manufacturing facility in informal text. This information is key to improving maintenance decisions, such as scheduling, dispatching, diagnosis, and inventory management, but is difficult to access due to the informal and domain specific nature of the text. Methods are available to aid manufacturers with parsing through this information, however small-to-medium sized manufacturers (SMMs) still have issues in implementing NLP solutions in practice. To this end, this paper discusses lessons learned in applying a NIST developed methodology to SMMs maintenance data.

1 Introduction

Within a manufacturing facility, maintenance logs that capture repair information, e.g., the problem, the solution, or the cause, are often completed by various operators or maintenance technicians. These technicians and operators often do not follow a set terminology or structure when entering this information. These inconsistencies in entering data even occur when only one person captures such information, such as when a manager enters all maintenance logs into a database. Due to such data logging inconsistencies, it is often difficult to observe or discover patterns or actionable information, particularly when a supervisor that is not directly involved with the maintenance process is reviewing the maintenance logs.

NIST researchers have developed technology using text analytics that has the ability to address this deficiency through its ability to assign tags, identify patterns, and extract actionable information from industrial data logs (Sexton Nestor, 2019). This methodology and subsequent analysis techniques have been developed for some time (Sexton, 2018; Sharp, 2019; Brundage, 2018; Sexton, 2017, Sharp, 2017). The software is open-source and available on GitHub[†] for all to use. Currently, the software helps maintainers annotate their Maintenance Work Order (MWO) data through a process called "tagging". The MWOs are inputted as comma-separated variable (.csv) files with UTF-8 encoding into the Nestor GUI and the user goes through the tagging process to create an annotated, tagged MWO dataset.

[*] Corresponding Author: michael.brundage@nist.gov
[†] https://www.nist.gov/services-resources/software/nestor

Brundage, Michael; Pavel, Radu. "Using Text Analytics Solutions with Small to Medium Sized Manufacturers: Lessons Learned." Paper presented at Model-Based Enterprise Summit, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020.

Considering the potential of the technology, NIST works with industry to further refine and improve their solution through assessment trials with data from manufacturers that can help reveal opportunities for improvement both in technology efficiency and in robustness of its applicability. TechSolve is working with NIST to assess the capabilities of the technology using maintenance data from manufacturing organizations willing to learn more about the potential advantages and suitability of such technology for the annotation, organization and analysis of their maintenance work orders/logs.

2 Data Collection Process

TechSolve leveraged its network of manufacturers to identify and recruit companies considered good candidates for this effort. A list of companies was compiled and readied for engagements starting in January 2019. Twenty seven (27) companies were contacted and assessed. Due to confidentiality constraints, the name of the companies cannot be disclosed. However, a list of their NACIS (North American Industry Classification System) codes and main characteristics is provided in Table 1, below.

Table 1. NAICS code, approximate number of employees, and the annual revenue for the companies contacted during this project (companies listed in random order)

No	**NAICS Code**	**Employees**	**Annual Sales**	**Notes**
Company 1	332119 - Metal Crown, Closure, and Other Metal Stamping (except Automotive)	60	$19M	Provided data
Company 2	336350 - Motor Vehicle Transmission and Power Train Parts Mfg	189	$37M	Provided data
Company 3	326199 - All Other Plastics Product Mfg 333514 - Special Die and Tool, Die Set, Jig, and Fixture Mfg	42	$10M	Declined to provide data

Company 4	332111 - Iron and Steel Forging	16	$4.2M	Declined to provide data
Company 5	332710 - Machine Shops	30	$629K	Declined to provide data
Company 6	333111 - Farm Machinery and Equipment Mfg	72	$42M	Provided data
Company 7	442299 - All Other Home Furnishings Stores	10	$1.5M	No electronic files
Company 8	334413 - Semiconductor and Related Device Mfg	142	$48M	Concerned with trade secrets/confidentiality
Company 9	311612 - Meat Processed from Carcasses	360	$25M	Expressed interest but no follow-up from company
Company 10	332710 - Machine Shops (Primary)	30	$6.1M	No follow-up from company
Company 11	322211 - Corrugated and Solid Fiber Box Manufacturing (Primary)	15	$5.9M	Declined due to limited availability of data
Company 12	333413 - Industrial and Commercial Fan and Blower and Air Purification Equipment Manufacturing (Primary)	31	N/A	Did not express interest in the opportunity

Company 13	335999 - All Other Miscellaneous Electrical Equipment and Component Manufacturing (Primary)	127	$25M	Management did not consider they have significant equipment and associated maintenance to qualify for this project
Company 14	423830 - Industrial Machinery and Equipment Merchant Wholesalers (Primary)	200	$1.7M	Management did not want to pursue opportunity
Company 15	333618 - Other Engine Equipment Manufacturing (Primary)	13	N/A	Did not express interest in the opportunity
Company 16	333249 - Other Industrial Machinery Manufacturing (Primary)	6	N/A	Management admitted they do not yet collect data in electronic format
Company 17	336390 - Other Motor Vehicle Parts Manufacturing (Primary)	2	N/A	Too small; limited maintenance necessary
Company 18	332710 - Machine Shops (Primary)	235	$39M	Very slow to reply. Too busy to commit for opportunity
Company 19	333922 - Conveyor and Conveying Equipment Manufacturing (Primary)	800	$800 M	Did not express interest in the opportunity

Company 20	331524 - Aluminum Foundries (except Die-Casting) (Primary)	18	$2M	Committed to send data but stopped short of sending a file
Company 21	332812 - Metal Coating, Engraving (except Jewelry and Silverware), and Allied Services to Manufacturers (Primary)	21	$2.8M	Did not express interest in the opportunity
Company 22	811310 - Commercial and Industrial Machinery and Equipment (except Automotive and Electronic) Repair and Maintenance (Primary) 336390 - Other Motor Vehicle Parts Manufacturing (Secondary)	366	$76M	Did not express interest in the opportunity
Company 23	336412 - Aircraft Engine and Engine Parts Manufacturing (Primary)	100	$16M	Did not express interest in the opportunity
Company 24	333511 - Industrial Mold Manufacturing (Primary)	183	N/A	Expressed interest but declined sending files
Company 25	334418 - Printed Circuit Assembly (Electronic Assembly) Manufacturing (Primary)	170	$44M	Did not express interest in the opportunity

Company 26	332911 - Industrial Valve Manufacturing (Primary)	150	$50M	Expressed interest but declined sending files
Company 27	333912 - Air and Gas Compressor Manufacturing (Primary)	50	$1.3M	Expressed interest but declined sending files

Although some of the companies expressed interest in the program, they withheld from sharing data over confidentiality and trade secret concerns. Other companies specified that they did not collect data, although they are interested to implement "best practices" and appropriate software solutions, such as computerized maintenance management systems (CMMS) or enterprise resource planning (ERP) systems. Such companies expressed the need for help in identifying those "best practices" and appropriate software, and mentioned that they had difficulties identifying a solution suited for them due to lack of knowledge in the field. In other cases, the companies were collecting maintenance information but could only output it in a printed form and were unable to export files in excel or .csv file format.

3 Lessons Learned

The companies collecting data in electronic format typically used three types of software: 1) a non-maintenance specific database (e.g., access or excel), 2) a computerized maintenance management system (CMMS) (e.g., Fiix), or 3) more generic planning system, such as the Enterprise Resource Planning (ERP) system (e.g., Plex). The companies looking to upgrade their maintenance work order capturing routine to an electronic platform, expressed interest in best practices and available/recommended solutions on the market – e.g. what would be the criteria to choose a good system for us? What system would be best for us? The companies that were already in possession of a software platform were interested to know if their practices and the way they are collecting the information are aligned with best practices. In addition, the manufacturers were interested in what would be more efficient and relevant analytics and charting for the maintenance work order they collect.

Concerning the engagement with industry, it was found that manufacturers recognize that the health and maintenance of the manufacturing assets represent an important area of their operations. The importance of maintenance appeared to be directly proportional with the size of the company and the cost of the product being manufactured. Nevertheless, from the contacted manufacturers that were engaged in communications with TechSolve, approximately 30% did not seem to have computerized means of capturing the maintenance work order data. This has been justified either from the perspective of the size of the company (too small), or the limited complexity of the equipment (e.g. conveyors or welding equipment). A limited number of companies, approximately three, confirmed they are collecting maintenance work order data and expressed interest in providing data; however, they were unable to export the data in csv or excel format. The main findings of the interactions with the companies that were contacted are summarized below:

- Approximately 75% of the companies that were contacted for this initiative expressed interest to learn more. However, only 50% of the interested companies moved forward with phone conversations or in-person visits. Eventually, from all contacted companies, only seven expressed intent to provide files, of which only three provided files eventually. From the three organizations that provided files, one was very concerned with the confidentiality of the data to the point that all operator names, asset names, and their locations had to be coded/changed.

- The majority of the companies expressed concern with and asked for maintaining the confidentiality of the information. If data sharing would be desired and further publication of the results, then the data should be stripped of identifiers and the provider should approve its release before the publication of the data and/or of the results.
- The companies compliant with ISO 9001 and AS9100 were more likely to have maintenance work order data.
- The companies that have maintenance records typically use a CMMS or ERP system to capture the information, and the work orders are logged in a database.
- Concerning the use of maintenance work order data, some companies appeared to have software with various capabilities to generate graphs or run statistics. However, the full functionality of the software, or the actual use of the software capabilities was not presented to the TechSolve team. Nevertheless, all companies expressed the desire to get better analytics and ways of visualizing data that would allow them to better understand the maintenance activities and extract actionable information.
- Long term, the companies expressed interest in implementing monitoring systems that would enable condition-based maintenance approach, versus reactive or preventative approaches.
- With regard to the maintenance work order data, all companies expressed interest in a solution that would help them better organize that data, and were interested to learn more about NIST's efforts on guidelines, standardization, and technology development addressing the manufacturing assets maintenance
- Due to the variety of systems used to collect data, the files shown had various column headers. Although only a limited number of files have been provided, the sample covers the typical scenarios discussed with industrials that span from custom made spreadsheets with small number of columns in an Excel file or Access database or using CMMS files with very large number of columns
- The variety of data collection format or the confidentiality restrictions imposed initial organization and filtering of the data files to enable proper processing with the NESTOR software and sharing the information with the NIST team.

After the examination of the maintenance work order files provided by the manufacturers, the following observations became apparent:

- Each company seem to collect data in its own, custom way, based on internal needs and guidance from the software provider; however, no particular "best practices" were pointed out or noticed.
- The names of the columns describing the maintenance task and/or resolution was different across the processed files.
- There often are no accurate records of the actual time it took to repair one item.
- There is limited information of who noticed the fault and who repaired the fault.
- Some descriptions are too simplistic, others may only be understood by someone that is very familiar with the manufacturing asset.
- It is needs to be clear to what extent the information is used for analysis and potential improvement opportunities. In general, the users would like to be able to derive (with simplicity) additional analytics/charting facilitating actionable information.

4 Conclusions and Future Work

This paper discusses lessons learned with SMMs for implementing a text analytics solution for analyzing maintenance work orders. The biggest concern of the manufacturers was providing proprietary information for analysis, thus, anonymization methods are important to improve the overall text analytics process. Most manufacturers that collected data had analytics and visualizations, but wanted more intuitive tools. Lastly, these manufacturers expressed interest at more predictive capabilities for discovering maintenance needs in the future and overall, the manufacturers involved in this initiative expressed interest in efforts associated with PHM for manufacturing assets.

The manufacturers involved in this study agreed that it would be very helpful to have guidance on best practices and product selection criteria with regard to capturing and processing maintenance work order data. Standards in the space of text analytics for manufacturers are needed to aid manufacturers in performing this analysis themselves (Weiss, 2019; Sexton Standards, 2019). The natural language processing concept and the availability of a technology to be used for organizing and annotating their data was regarded positively.

NIST Disclaimer

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

References

Sexton, Thurston B., and Michael P. Brundage. Nestor: A Tool for Natural Language Annotation of Short Texts. No. Journal of Research of the National Institute of Standards and Technology. 2019.

Sexton, Thurston, et al. "Benchmarking for keyword extraction methodologies in maintenance work orders." Proceedings of the Annual Conference of the PHM Society. Vol. 10. No. 1. 2018.

Sharp, Michael, et al. "Selecting Optimal Data for Creating Informed Maintenance Decisions in a Manufacturing Environment." Model-Based Enterprise Summit 2019. 2019.

Brundage, Michael P., et al. "Developing maintenance key performance indicators from maintenance work order data." ASME 2018 13th International Manufacturing Science and Engineering Conference. American Society of Mechanical Engineers Digital Collection, 2018.

Sexton, Thurston, et al. "Hybrid datafication of maintenance logs from ai-assisted human tags." 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017.

Sharp, Michael, Thurston Sexton, and Michael P. Brundage. "Toward semi-autonomous information." IFIP International Conference on Advances in Production Management Systems. Springer, Cham, 2017.

Weiss, Brian A., et al. Summary Report on the Industry Forum for Monitoring, Diagnostics, and Prognostics for Manufacturing Operations. No. Advanced Manufacturing Series (NIST AMS)-100-23. 2019.

Sexton, Thurston, and Michael P. Brundage. "Standards Needs for Maintenance Work Order Analysis in Manufacturing." Model-Based Enterprise Summit 2019. 2019.

Using Text Visualization to Aid Analysis of Machine Maintenance Logs

Senthil Chandrasegaran[1], Xiaoyu Zhang[1], Michael P. Brundage[2], and Kwan-Liu Ma[1]

[1] Department of Computer Science, University of California, Davis, Davis, CA
{schandrasegaran, xybzhang, klma}@ucdavis.edu
[2] National Institute of Standards and Technology, Gaithersburg, MD
michael.brundage@nist.gov

Abstract

Maintenance and error logs for machines in manufacturing organizations are typically written as informal notes by operators or technicians working on the machines. These logs are written using a combination of common language and internally-used abbreviations and jargon. Due to inconsistencies in the terminology used during error logging and in identifying root causes of issues, the data needs to be cleaned before automated analyses can be effectively used. This can require a human to go through and clean/tag the data, disambiguate multiple terms, and sometimes assign additional tags to the data objects to aid automated classification. With some organizations storing over a million records of legacy maintenance report data, this is not entirely feasible. We introduce a visual analytic approach to help analysts sift through such heterogeneous datasets so that the inconsistent data can be tagged and categorized with minimum manual effort. Though such data typically includes metadata such as date, time, severity, machine IDs, etc., in this paper we focus on the manually-entered text descriptions. We use metrics such as word occurrence frequency and information-theoretic metrics to visually highlight common and uncommon issues and fixes that occur in the maintenance logs. We illustrate our approach with data from industry and discuss future research directions to address scalability, metadata, and other approaches for grouping similar logs.

1 Introduction

Machine error diagnosis and prediction is an issue in which the manufacturing industry heavily invests due to its direct effect on machine availability and throughput. Some organizations often maintain cross-functional teams of engineers with expertise covering design, analysis, and manufacturing to help identify and correct such problems quickly. With the growth of Smart Manufacturing and inexpensive and easy to use sensors, the demand for data-driven solutions for machine diagnostics and prognostics has increased. Organizations thus maintain machine maintenance and error logs to help analysts identify patterns and subsequently formalize root-cause analysis of errors. This, in turn, helps organizations plan preventive and predictive maintenance practices.

However, maintenance and error logs for machines can be both human- and sensor-generated. Human-generated logs are typically written as informal notes by technicians, who often use their own jargon when referring to machines, parts, and processes. These terms are frequently not consistent across groups in the organization, making it difficult for analysts to identify similar logs. Sensor-generated logs tend to be highly general and lack relevant and contextual machine-specific information. This data poses the opposite problem: sensors often have similar logs even though the logs are generated from a variety of machines for a variety of problems. Finally, both

Brundage, Michael; Chandrasegaran, Senthil; Zhang, Xiaoyu; Ma, Kwan-Liu. "Using Text Visualization to aid Analysis of Machine Maintenance Logs." Paper presented at Model-Based Enterprise Summit, Gaithersburg, MD, US. March 30, 2020 - April 03, 2020.

human- and machine-generated logs are sometimes inconsistent: humans are prone to errors in identification and labeling of symptoms and diagnoses, while sensors can have errors that may result in erroneous/missing logs or corrupted data.

There are ongoing efforts to clean, consolidate, tag, and categorize maintenance and error logs to aid automated diagnostics and prognostics. This cleaning effort requires manual tagging and repairing of data, disambiguation of terms, and assigning of specific terms to aid automated classification of the data. With some organizations storing over a million records of legacy maintenance report data, this is not entirely feasible without aid of tools. Recent semi-automated approaches have used the human in the loop along with natural language processing techniques to aid the above disambiguation and tagging. However, these methods still require making assumptions in the process of cleaning and categorizing data to extract useful and/or actionable information.

To aid human analysts in viewing large datasets, grouping them, and observing patterns and anomalies that aid labeling and categorization of data, we propose the use of visual analytics. The science of visual analytics supports data analysis using computational techniques and interactive visualizations [7]. Specifically, it allows analysts to forage for information, collect evidence, and form schema that leads to hypotheses, a process called the visual sensemaking loop [20]. In this paper, we introduce a visual analytics approach meant for aiding qualitative text analysis and categorization, and apply it to the analysis and categorization of machine log data. We focus on the manually-entered text descriptions and outline requirements that need to be fulfilled to manually analyze and tag such log data for better sensemaking. We describe how the visual analytic approach addresses these requirements, and illustrate the approach with a use-case scenario of maintenance log data from the industry. We close with recommendations for incorporating metadata and approaches for better scalability.

2 Background

With increasing emphasis on smart manufacturing and a push toward eliminating machine downtime, process monitoring, diagnostics, and prognostics have gained prevalence. The complexity and volume of data that needs to be sifted through to achieve this improved maintenance of equipment have prompted the application of visual analytics into product lifecycle management (PLM) [21]. This potential application area was anticipated almost at the inception of visual analytics when Keim et al. [17] suggested that visual analytics may be used in engineering for analyzing complex data that arise from design, production, and feedback from product use. In this section, we will provide a background on the complexity of making sense of machine error and maintenance logs, and the application of visual analytics to address this complexity.

2.1 Processing Human- and Sensor-Generated Logs

System log analysis—analysis of logs automatically generated by the system—is commonly used to track system resilience. It is also used in the case of failure for root cause analysis and in the case of preventive maintenance to identify recurring patterns, such as temporal, systemic, or even seasonal. Typical tools used for such analyses use automated analyses and seldom resort to visualization approaches.

Automatic log analysis tools [8, 13, 14, 16, 19, 30] typically use a range of analyses such as correlation analysis, signal analysis, pattern mining, correlations, resilience analyses at the application level, and spatial/temporal event analysis. For instance, HELO (Hierarchical Event Log Organizer) [14], an event log mining tool, extracts event formats by pattern-mining log files from large-scale supercomputers, using predefined message templates. A model-based approach is used by ELSA (Event Log Signal Analyzer) [13], a toolkit for event prediction. It models the normal flow at a stable event state, and in the event of system failure, tracks the abnormal flow of events using a combination of data mining and signal processing.

There exist visualization-oriented tools for tracking and analyzing machine logs, but these are few, and most of them use relatively basic visualizations. For instance, LogMaster [12] and LogAider [8] use generic visualizations for mining event correlations. LogAider reveals potential correlations that include across-field (through probabilistic analysis of fields), spatial, and temporal correlations. LogMine [16] is a framework for the unsupervised, scalable end-to-end one-pass analysis of large-scale, heterogeneous logs. LogDiver [19] supports lossless data compression, models application failure paths, and cross-validates models and/or results of analyses. More recently, machine learning approaches such as DeepLog [10] have been introduced. Specifically, Deeplog uses a deep neural network model which uses stacked Long Short-Term Memory (LSTM) to detect anomalies, and dynamically updates the models to accommodate for changing log patterns.

The idea of using visualization and visual analytics for monitoring and diagnostics in factories is a relatively new research area. Recent work includes ViDX [29], a visual analytic system for historical analysis and real-time monitoring of factory assembly lines. ViDX uses visualization principles to create outlier-aware aggregate representations of process data and employs user-steerable algorithms for outlier detection. La VALSE [15] is a scalable log visualization tool that uses multiple visualizations for interactive event analysis based on multiple logs. ViBR [5] is a system that visualizes bipartite relationships using a minimum description length principle to aggregate the relationships. The system has been successful in log analyses that include vehicle fault diagnostics by identifying co-occurring faults, comparing faults that co-occur in different vehicle clusters, and comparing faults across vehicles with shared properties.

2.2 Visual Analytics for Text Data

Root cause analysis and preventative action is a crucial area of interest to the manufacturing industry, necessitating logging maintenance and error log data, as discussed earlier. Approaches to parse this data for an automated or even semi-automated solution for diagnosis or prognosis has typically involved knowledge bases [4], manual "tagging" systems assisted by natural-language text parsing support [23][1], and information extraction methods applied to maintenance logs [24, 25]. While our approach also proposes the use of natural language processing (NLP) techniques, we use visual analytics to keep the human in the loop for correcting and tagging the parsed data through the visual representation of and interaction with the data processing results.

Defined as "the science of analytical reasoning facilitated by interactive visual interfaces" [7, p. 4], visual analytics uses visualization support throughout the *process* of analyzing (typically unstructured) datasets. In other words, visual analytics makes "*our way of processing* data and information transparent for an analytic discourse." [17, p. 155]. At the center of all visual analytic systems is the analyst—the human in the loop—who is aided by the system in combining complex datasets, collect evidence, identify correlations, and develop insights. At every stage of this process, the analyst is aided by a combination of visualizations and algorithms.

Visual analytics support for text analysis often focuses on analyzing connections between multiple sources of text, from intelligence reports to news articles to even unstructured social

[1] An open source tool for this process, called Nestor is available here: https://www.nist.gov/services-resources/software/nestor

media texts such as tweets and posts on forums. Some of the earliest text analytic tools were designed for intelligence analysis. Of these, Jigsaw [26] is one of the more prominent and still-used tools. It identifies connections between documents using entities in text data and metadata, highlights these connections to the user, and allows the user to reorganize this information to aid their insight-gathering process. It uses coordinated views such as graphs, calendars, and document overviews, all of which can be filtered and edited by the analyst to identify potential security threats. Other approaches make more use of metadata, such as time-stamps. For instance, Tiara [28], a system for temporal analysis of text documents, is used to analyze data relevant to emails, instant messages, and even patient records. It uses statistical text analysis techniques such as topic modeling to categorize the document collections thematically based on their content, and shows the variation of themes over time. It also allows users to select and examine any theme-based collection in detail, across and at defined time intervals. Other topic-modeling-based text analysis tools include HierarchicalTopics [9], which as the name suggests, uses a hierarchical topic modeling algorithm to identify themes within themes. It combines this with a temporal view showing the evolution of topics over time and allows users to explore and edit topics hierarchically. Other approaches are more suitable for single or very few, but large documents such as historical texts. An example is VariFocal Reader [18], which uses automated annotations and topic modeling to reveal thematic and structural patterns that are useful when analyzing large documents.

In this paper, we adapt our prior work that uses visual analytics with a dominant text visualization component that we developed to aid qualitative analysis of text data [6]. We do this by helping the user identify concepts of interest, categorize associated text, and use their custom categorizations to further analyze the text. We illustrate the suitability of this approach in helping users identify patterns and inconsistencies in any terminology used in machine logs. This will help analysts create useful categorizations of machine logs that will help problem diagnosis, and to subsequently create machine learning models.

3 Design

While individual fields of machine maintenance logs may vary between organizations, they usually have some common features, such as the machine identifier, problem description, the description of the remedial action taken, and the dates on which the problem was reported and closed. While it is possible to “group” these logs by some of the features such as machine ID, it becomes less obvious to group the logs based on the type of problem, the type of solution, or patterns in the dates on which they tend to occur. Such categorization often requires the expertise and insight borne by experience. The goal of our approach is to help such experienced personnel sift through and examine large volumes of data without needing to examine each record closely.

3.1 Design Rationale

We draw from research in visual analytics—“the science of analytical reasoning facilitated by interactive visual interfaces” [7, p. 421]—to design an appropriate interface for our approach. We identify the following requirements for prognostics of machine maintenance.

R1 Identify Common Occurrences: One of the main requirements in machine maintenance log analysis is to identify recurring problems that—while individually may not cause significant downtime—through their frequency of occurrence cost significant resources in repairing and downtime. These may not always be linked to the same kinds of machines, or even have the same descriptions.

R2 Identify Patterns in Occurrences: Some maintenance issues may manifest as several problems that occur together or in succession to cause a much more significant issue than the individual reports suggest. Other issues may occur only in some kinds of machines, or when some operators are working certain machines, or even certain days of the week, month, or year. Combined with the earlier-identified issue of inconsistency in the descriptions, the need to identify patterns in maintenance logs is only matched by the challenges posed in identifying such patterns.

R3 Identify Anomalies: When taking stock of problems that occur over a long period, there may be a need to identify rare, yet significant problems. These could refer to the problems themselves, or their rare occurrence in a specific machine or part. Such anomalies could be lost to cursory scrutiny when looking for commonly-occurring problems, but if ignored could escalate over time.

R4 Allow Manual Categorization: Identifying patterns, anomalies, and common issues is often not a single-stage process. The relevant analyst or domain expert may need to tag certain groups of problems with a descriptor, add a memo for continued monitoring, or even need verification from a colleague.

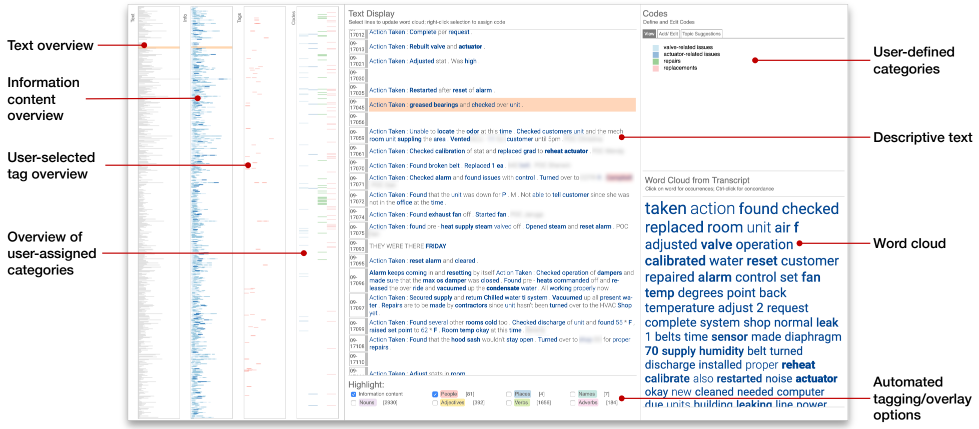
R5 Aid Iterative Analyses: Once manual categories are identified, the system should allow the user to filter the existing data using these categories, which will further reveal commonly-occurring keywords.

3.2 Interface Design

Based on the above rationale, we decided on a primarily text-based visualization approach, shown in Fig.1. The visualization is largely extended from our prior work on developing a visual analytic approach to aid qualitative text analysis [6]. Since our focus in this work is on the content (and less so on the metadata), the text component of the data is shown in the central panel. These descriptions are logged by the machine technicians and/or operators and include reports that can describe the problem, the solution, or both. This being a preliminary approach that examines how the existing qualitative analysis system can be used for analyzing patterns in the data, we do not incorporate temporal or other metadata such as machine IDs, operator IDS, severity or cost-related information.

The text shown in the central column follows a "skim formatting" [3] where the font weight for each word corresponds to a predefined criterion. In our case, we use the word information content [22], which is based on the assumption that the less frequently-used a word is in a corpus, the more information it contains. For a more focused application, we can use analyst-defined metrics that give greater weight to keywords associated with rare and severe issues (requirement **R3**) in conjunction with—or instead of—such generic metrics. The information content metric can also be computed on a specific domain, such as existing corpora of operation or repair manuals.

On the right is a word cloud that is automatically computed from the uploaded text. It is scaled proportional to frequency and the words are arranged sequentially in descending order of frequency. The skim formatting described earlier is applied to the word cloud as well. Selecting a set of text in the central column will filter the word cloud to reflect only the selection. The word


Text overview
Information content overview
User-selected tag overview
Overview of user-assigned categories
Text
Info
Tags
Codes
User-defined categories
Descriptive text
Word cloud
Automated tagging/overlay options
View
Add/ Edit
Topic Suggestions

Text Display
Select lines to update word cloud; right-click selection to assign code

09-17012 Action Taken : Complete per request .
09-17013 Action Taken : Rebuilt valve and actuator .
09-17021 Action Taken : Adjusted stat . Was high .
09-17030
09-17035 Action Taken : Restarted after reset of alarm .
09-17045 Action Taken : greased bearings and checked over unit .
09-17056
09-17059 Action Taken : Unable to locate the odor at this time . Checked customers unit and the mech room unit suppling the area . Vented [redacted] customer until 5pm . [redacted]
09-17061 Action Taken : Checked calibration of stat and replaced grad to reheat actuator . [redacted]
09-17070 Action Taken : Found broken belt . Replaced 1 ea . [redacted]
09-17071 Action Taken : Checked alarm and found issues with control . Turned over to [redacted]
09-17072 Action Taken : Found that the unit was down for P . M . Not able to tell customer since she was not in the office at the time .
09-17074 Action Taken : Found exhaust fan off . Started fan . [redacted]
09-17075 Action Taken : found pre - heat supply steam valved off . Opened steam and reset alarm . POC [redacted]
09-17093 THEY WERE THERE FRIDAY
09-17095 Action Taken : reset alarm and cleared .
09-17096 Alarm keeps coming in and resetting by itself Action Taken : Checked operation of dampers and made sure that the max os damper was closed . Found pre - heats commanded off and re-leased the over ride and vacuumed up the condensate water . All working properly now .
09-17097 Action Taken : Secured supply and return Chilled water ti system . Vacuumed up all present wa-ter . Repairs are to be made by contractors since unit hasn't been turned over to the HVAC Shop yet .
09-17099 Action Taken : Found several other rooms cold too . Checked discharge of unit and found 55 * F , raised set point to 62 * F . Room temp okay at this time . [redacted]
09-17108 Action Taken : Found that the hood sash wouldn't stay open . Turned over to [redacted] for proper repairs .
09-17110
09- Action Taken : Adjust stats in room

Highlight:
☑ Information content ☑ People [81] □ Places [4] □ Names [7]
□ Nouns [2930] □ Adjectives [392] □ Verbs [1656] □ Adverbs [184]

Codes
Define and Edit Codes

- valve-related issues
- actuator-related issues
- repairs
- replacements

Word Cloud from Transcript
Click on word for occurrences; Ctrl-click for concordance

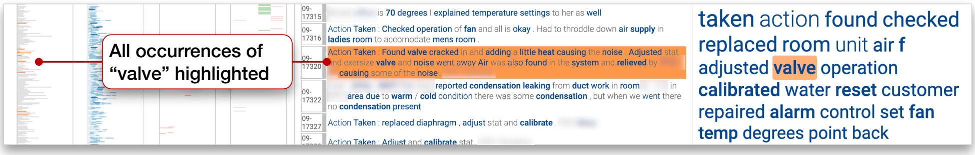
taken action found checked
replaced room unit air f
adjusted valve operation
calibrated water reset customer
repaired alarm control set fan
temp degrees point back
temperature adjust 2 request
complete system shop normal leak
1 belts time sensor made diaphragm
70 supply humidity belt turned
discharge installed proper reheat
calibrate also restarted noise actuator
okay new cleaned needed computer
due units building leaking line power

Figure 1: The interface adapted from our earlier work in Chandrasegaran et al. [6], shown here with approximately 600 records of maintenance logs for HVAC (heating, ventilation, and air-conditioning) systems at a specific site.

cloud can be used to identify commonly-occurring terms in the maintenance logs (requirement **R1**).

A set of checkboxes on the bottom are used to highlight parts of speech or named entities (person, place, names). These can be useful to highlight when the analyst is looking for logs that mention geographic locations, or when the names of operators/repair persons are mentioned.

An overview pane on the left shows several overview visualizations. The first is a "text overview" that simply gives a mini-map view of all the records in the collection. It also shows the position of the current record of interest (as an orange bar) corresponding to the text in the central panel on which the mouse currently hovers. As shown in Fig. 2, selecting a word in the word cloud shows all its occurrences in the main text as well as in the overview panel (requirement **R2**).

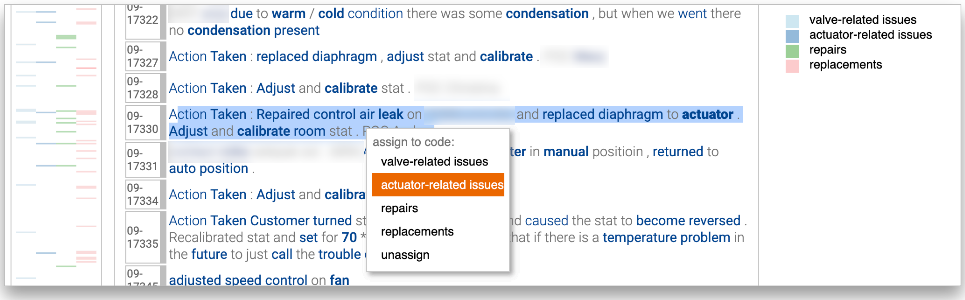


09-17315 [redacted] is 70 degrees I explained temperature settings to her as well
09-17316 Action Taken : Checked operation of fan and all is okay . Had to throddle down air supply in ladies room to accomodate mens room .
09-17320 Action Taken : Found valve cracked in and adding a little heat causing the noise . Adjusted stat and exersize valve and noise went away Air was also found in the system and relieved by [redacted] causing some of the noise . [redacted]
09-17322 [redacted] reported condensation leaking from duct work in room [redacted] in [redacted] area due to warm / cold condition there was some condensation , but when we went there no condensation present
09-17327 Action Taken : replaced diaphragm , adjust stat and calibrate . [redacted]
09- Action Taken : Adjust and calibrate stat [redacted]

taken action found checked
replaced room unit air f
adjusted valve operation
calibrated water reset customer
repaired alarm control set fan
temp degrees point back

Figure 2: A detail of the interface showing how selecting a word from the word cloud (right) highlights all records where that word occurs, in both the detailed text view and the text overview panes.

Additional overviews include an information content heatmap that provides an overview of the skim formatting described earlier, parts-of-speech/named entity tag overviews, and an overview of user-applied categories. These categories are specified in an input field on the top right part of the interface (Fig. 1). Once the categories are specified, they can be assigned to individual fields or groups of sequential fields as shown in Fig. 3 (requirement **R4**). Assigning a category to one or more fields of text updates the overview visualization immediately to the left of the text display. Once the categories of interest have all been assigned, co-occurring problems and—once temporal data can be integrated—temporal and recurring patterns can be visually identified, and these co-occurrences can be further tagged and newer categories assigned to them iteratively (requirement **R5**).



09-17322 [redacted] due to warm / cold condition there was some condensation , but when we went there no condensation present

09-17327 Action Taken : replaced diaphragm , adjust stat and calibrate . [redacted]

09-17328 Action Taken : Adjust and calibrate stat . [redacted]

09-17330 Action Taken : Repaired control air leak on [redacted] and replaced diaphragm to actuator . Adjust and calibrate room stat . [redacted]

09-17331 [redacted] ter in manual positioin , returned to auto position .

09-17334 Action Taken : Adjust and calibra

09-17335 Action Taken Customer turned st nd caused the stat to become reversed . Recalibrated stat and set for 70 * hat if there is a temperature problem in the future to just call the trouble

09-17345 adjusted speed control on fan

Figure 3: Detailed view of the text and code (category) definition and overview fields showing the categories assigned. A category is assigned by selecting a block of text and assigning a category from a drop-down menu shown above. When a category is assigned to a text, it updates the overview visualization on the left.

4 Implementation

As explained in Section 3, the system presented here is adapted from our earlier work directed at qualitative text analysis [6]. The system is implemented as a web-based application in HTML5 and JavaScript, with a Node.js backend where the data is uploaded and processed to be visualized on the browser. Most of the language processing operations, including tokenization, parts-of-speech tagging, named-entity recognition, and information content measurement are performed at the server end using Python's Natural Language Toolkit [1], and the Stanford POS [27] and NER [11] taggers. At the front end, the interactive visualizations are created using the D3.js [2] JavaScript library. The code is available as open-source[2].

5 Use Case Scenario

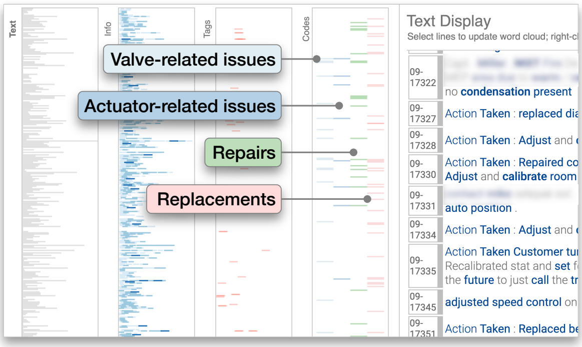
To illustrate the system in action, we present a use-case scenario with a dataset of 600 records concerning the maintenance of an HVAC (heating, ventilation, and air-conditioning) system of a set of office buildings. Since our focus is primarily on the text descriptions, we remove all temporal and machine/operator-related metadata before uploading it into our system. Refer back to Fig. 1 for an overview of this dataset when processed and viewed in the system.

We consider an analyst—a maintenance specialist interested in identifying commonly-occurring patterns where repair and/or replacement is required. Once the analyst loads the data, they take a closer look at the word cloud view (Fig. 1) and see that the more commonly-occurring terms seem to be generic terms—mostly verbs—that appear to be concerned with

[2] https://github.com/senthilchandrasegaran/textplorer/

remedial action, such as “taken”, “found”, “checked” etc. The letter “f” also appears frequently, and upon closer inspection, is revealed to be aggregated from all the mentions of temperatures in Fahrenheit. The first *item*—a component that finds frequent mention—is “valve”. Selecting this word in the cloud immediately highlights all its occurrences in the text and the overview (see Fig. 2). The highlights in the overview visualization show that “valve” does indeed appear fairly uniformly across the maintenance records. The analyst is curious if most of the valve-related issues also relate to actuators. They select “actuator” in the word cloud, but realize that they need to see co-occurrence patterns, i.e. cases where valve-related issues co-occur with actuator-related issues.

The analyst decides to create a category called “valve-related issues”, and another called “actuator-related issues”. They manually select every field that shows the occurrence of the word “valve” and assign the category “valve-related issues” to it. They follow a similar process for the actuators (as shown in Fig. 3). They are also curious to see the distribution of valves/actuators repaired and those that are replaced. They create two more categories called “repairs” and “replacements” and through a similar process, continue assigning categories.



Text Display
Select lines to update word cloud; right-cl

09-17322	[redacted] no condensation present
09-17327	Action Taken : replaced dia
09-17328	Action Taken : Adjust and
09-17330	Action Taken : Repaired co Adjust and calibrate room
09-17331	[redacted] auto position .
09-17334	Action Taken : Adjust and
09-17335	Action Taken Customer tu Recalibrated stat and set f the future to just call the tr
09-17345	adjusted speed control on
09-17351	Action Taken : Replaced be

Figure 4: Detail view of the categorization overview showing co-occurrences of the manually-created categories.

Once the categories are assigned, they inspect the co-occurrences of these categories closely (see overview in Fig. 1 and detailed view in Fig. 4). A close inspection of the co-occurrences shows the analyst that most valve- and actuator-related issues are repaired (with a few replacements), and that there are few cases where valve-related problems co-occur with actuator-related problems. The analyst continues with more inclusive terms for repair such as “fixed”, “removed”, “cleaned” and so on, assigning the same category of “repair” to them, to hunt for more patterns.

6 Conclusion

Visual analytics has been shown to be the best solution for sense-making when it comes to semi-structured data such as maintenance logs. In this paper, we illustrate how a visual analytics approach that was designed for qualitative text analysis can be used for analyzing the raw text from machine maintenance logs. Specifically, we identified requirements such as identifying common occurrences, patterns and anomalies, and the need for manual categorization and iterative analyses that an analysis tool should address for use in machine maintenance logs. We made the argument for how a visual analytic approach—which combines automated analysis techniques with human-in-the-loop interfaces—is suitable to address such requirements. We described our interface and with a use-case scenario, illustrated how the system can be used to identify similar maintenance logs, manually assign categories to these entries, and use category co-occurrence to form further insights.

Our current approach was illustrated with a machine log dataset with around 600 records. For our future work, we plan to extend our approach to be more scalable, as machine logs can extend to thousands or even millions of records. While visual representations such as word clouds and information-content maps are scalable, text overview and detail displays need to be redesigned to scale to such large records. One approach we plan to use is to incorporate machine log metadata to separate logs that may be unrelated (and can thus be examined separately). We also plan to use dimensionality-reduction techniques that can make use of metadata to automatically suggest clusters based on similarity metrics, or weights that can be derived through discussions with analysts. We will iteratively refine our approach through longitudinal studies with technicians experienced in machine maintenance for better environmental validity.

NIST Disclaimer

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

References

[1] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[2] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[3] Richard Brath and Ebad Banissi. Using text in visualizations for micro/macro readings. In *Proceedings of the ACM Intelligent User Interfaces Workshop on Visual Text Analytics*, 2015.

[4] Michael P Brundage, Boonserm Kulvantunyou, Toyosi Ademujimi, and Badarinath Rakshith. Smart manufacturing through a framework for a knowledge-based diagnosis system. In *Proceedings of the ASME International Manufacturing Science and Engineering Conference*, 2017.

[5] Gromit Yeuk-Yin Chan, Panpan Xu, Zeng Dai, and Liu Ren. ViBR: Visualizing bipartite relations at scale with the minimum description length principle. *IEEE transactions on visualization and computer graphics*, 25(1):321–330, 2019.

[6] Senthil Chandrasegaran, Sriram Karthik Badam, Lorraine Kisselburgh, Karthik Ramani, and Niklas Elmqvist. Integrating visual analytics support for grounded theory practice in qualitative text analysis. *Computer Graphics Forum*, 36(3):201–212, 2017.

[7] Kristin A Cook and James J Thomas. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005.

[8] Sheng Di, Rinku Gupta, Marc Snir, Eric Pershey, and Franck Cappello. Logaider: A tool for mining potential correlations of hpc log events. In *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pages 442–451, 2017.

[9] Wenwen Dou, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2002–2011, 2013.

[10] Min Du, Feifei Li, Guineng Zheng, and Vivek Srikumar. Deeplog: Anomaly detection and diagnosis from system logs through deep learning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1285–1298, 2017.

[11] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the Annual Meeting on Association for Computational Linguistics*, pages 363–370, 2005.

[12] Xiaoyu Fu, Rui Ren, Jianfeng Zhan, Wei Zhou, Zhen Jia, and Gang Lu. Logmaster: Mining event correlations in logs of large-scale cluster systems. In *IEEE Symposium on Reliable Distributed Systems*, pages 71–80, 2012.

[13] Ana Gainaru, Franck Cappello, and William Kramer. Taming of the shrew: Modeling the normal and faulty behaviour of large-scale HPC systems. In *IEEE International Parallel and Distributed Processing Symposium*, pages 1168–1179, 2012.

[14] Ana Gainaru, Franck Cappello, Stefan Trausan-Matu, and Bill Kramer. Event log mining tool for large scale hpc systems. In *European Conference on Parallel Processing*, pages 52–64, 2011.

[15] Hanqi Guo, Sheng Di, Rinku Gupta, Tom Peterka, and Franck Cappello. La VALSE: Scalable log visualization for fault characterization in supercomputers. In *EGPGV*, pages 91–100, 2018.

[16] Hossein Hamooni, Biplob Debnath, Jianwu Xu, Hui Zhang, Guofei Jiang, and Abdullah Mueen. Logmine: Fast pattern recognition for log analytics. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 1573–1582, 2016.

[17] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Information visualization*, pages 154–175. Springer, 2008.

[18] Steffen Koch, Markus John, Michael Wörner, Andreas Müller, and Thomas Ertl. Varifocalreader—in-depth visual analysis of large text documents. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1723–1732, 2014.

[19] Catello Di Martino, Saurabh Jha, William Kramer, Zbigniew Kalbarczyk, and Ravishankar K Iyer. Logdiver: A tool for measuring resilience of extreme-scale systems and applications. In *Proceedings of the ACM Workshop on Fault Tolerance for HPC at eXtreme Scale*, pages 11–18, 2015.

[20] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.

[21] Devarajan Ramanujan, William Z Bernstein, Senthil K Chandrasegaran, and Karthik Ramani. Visual analytics tools for sustainable lifecycle design: Current status, challenges, and future opportunities. *Journal of Mechanical Design*, 139(11):111415, 2017.

[22] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 448–453, 1995.

[23] Thurston Sexton, Michael P Brundage, Michael Hoffman, and Katherine C Morris. Hybrid datafication of maintenance logs from ai-assisted human tags. In *IEEE International Conference on Big Data*, pages 1769–1777, 2017.

[24] Thurston Sexton, Melinda Hodkiewicz, Michael P Brundage, and Thomas Smoker. Benchmarking for keyword extraction methodologies in maintenance work orders. In *PHM society conference*, volume 10, 2018.

[25] Michael Sharp, Thurston Sexton, and Michael P Brundage. Toward semi-autonomous information extraction for unstructured maintenance data in root cause analysis. In *IFIP International Conference on Advances in Production Management Systems*, pages 425–432, 2017.

[26] John Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[27] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*

Technology, pages 173–180, 2003.

[28] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, pages 153–162, 2010.

[29] Panpan Xu, Honghui Mei, Liu Ren, and Wei Chen. Vidx: Visual diagnostics of assembly line performance in smart factories. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):291–300, 2016.

[30] Ziming Zheng, Zhiling Lan, Byung H Park, and Al Geist. System log pre-processing to improve failure prediction. In *IEEE/IFIP International Conference on Dependable Systems & Networks*, pages 572–577, 2009.

Record Fast Polarization Switching Observed in Ferroelectric Hafnium Oxide Crossbar Arrays

Xiao Lyu[1], Mengwei Si[1], Pragya R. Shrestha[2,3], Jason P. Campbell[3], Kin P. Cheung[3] and Peide D. Ye[1,*]
[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA. *Email: yep@purdue.edu
[2]Theiss Research, La Jolla, USA. [3]National Institute of Standards and Technology, Gaithersburg, USA.

Abstract

The polarization switching speed of ferroelectric (FE) hafnium zirconium oxide (HZO) is studied with the device size down to sub-μm in lateral dimension. Ultrafast measurement of transient switching current on metal-ferroelectric-metal (MFM) device with a crossbar array or a single crossbar structure is performed to analyze the switching dynamics. A record fast polarization switching of 360 ps is achieved for 15 nm thick HZO with 0.1 μm^2 crossbar array device structure. The observed record switching speed is found to be limited by domain wall propagation speed in the nucleation limited switching process. It is further verified after significant reduction of RC delay of the devices and the implementation of crossbar array structure.

Introduction

Ferroelectric hafnium oxide [1] has been widely studied as a promising CMOS-compatible material in commercial ferroelectric device applications. The remarkable endurance and retention performance [2] make FE HfO$_2$ suitable for non-volatile ferroelectric random-access memory (FeRAM) and ferroelectric FET (FeFET) devices. However, fast operational speed [3-11] is crucial for FE HfO$_2$-based memory to replace the current commercial memory products. Previous works [3-5] demonstrated that switching speed of FE HfO$_2$ can reach sub-ns regime in a single MFM device with an area of ~6 μm^2. Direct fast speed measurement on a much smaller size MFM capacitor is challenging due to small capacitance and high access resistance of metal contacts.

In this work, we fabricated crossbar array devices with 100-200 nm in size along with single crossbar devices with similar total areas for comparison. We performed transient current measurements and analysis with an ultrafast pulse measurement setup. Polarization switching dynamics in single and array crossbar devices is studied and quantitatively characterized. By applying nucleation limited switching (NLS) model [12] to the net switching current, a record fast switching time of 360 ps is obtained from an array of 0.1 μm^2 crossbar device. It confirms experimentally that the domain wall propagation is the limiting factor for the measured hundreds of ps polarization switch speed instead of the RC delay.

Experiments

The devices were fabricated on an insulating sapphire substrate to exclude the impact of parasitic capacitance from the substrate in fast pulse measurements. The fabrication process is described in Fig. 1. Sputtered Au/W metal stack was used as bottom electrode to achieve small series resistance. Anisotropic wet etch of gold was performed after the dry etch of tungsten to prevent possible short circuit between the top and bottom electrodes. The bottom Ti/Au contact pads were formed by e-beam evaporation for a better contact resistance. The growth of 15 nm thick HZO was completed by atomic layer deposition (ALD) at 200 °C [3]. A 20 nm thick tungsten layer was sputtered after HZO deposition, followed by a rapid thermal annealing (RTA) in nitrogen at 500 °C for 1 min. The top Ti/Au electrodes were fabricated by a lift-off process. Dry etching of top W layer was performed using Ti/Au as the hard mask to realize device isolation.

In order to precisely measure the polarization current flowing through a very small size capacitor, crossbar array structure is used to collectively measure a large number of individual small capacitors at the same time. Crossbar array devices consist of ten 1 μm-wide metal electrode stripes at the bottom and twenty 100 or 200 nm-wide metal stripes on the top, as seen in Fig. 2(a). Single crossbar devices with large sizes but same total areas, as shown in Fig. 2(b) and Fig. (3), are designed as control devices for comparison. Fig. 4 shows the cross-section schematic and layer details of the MFM structure. All devices show strong ferroelectricity in polarization versus electric field (P-E) as illustrated in Fig. 5. Positive-up-negative-down pulse sequences, as shown in Fig. 6, are generated by an arbitrary pulse generator to perform direct measurement of transient current. The circuit diagram of the ultrafast pulse measurement setup and established methodology can be found in [3]. Signal reflection is efficiently suppressed by applying impedance matched probes, 50 Ohm terminations and the pick-off tee.

Results and Discussion

Fig. 7 shows a representative PUND transient current measurement. Both pulses share the same input voltage waveform with a 200 ps rise time and the corresponding transient current responses of the switching (first) pulse and non-switching (second) pulse are measured. The transient current I_{pulse1}, I_{pulse2} and the net switching current I_{FE} are plotted in Fig. 8 for a crossbar array device with 15 nm thick HZO as a representative case. The switching current is extracted by the subtraction of I_{pulse1} and I_{pulse2}. The time response of ferroelectric polarization switching charge can be determined by integrating the corresponding net switching current as seen in Fig.9. The FE polarization switching follows the NLS model, which can be described by $P = P_S(1 - \exp(-\left(\frac{t}{t_0}\right)^2))$ in thin film. The t_0 in NLS model is a characteristic switching time constant [10-11] with a record fast 360 ps on 0.1 μm^2 and 640 ps on 0.2μm^2 array devices with 15 nm HZO.

These newly fabricated crossbar devices have lower series resistance compared to unoptimized devices in [3]. Fig. 10 shows the comparison of (a) switching speed and (b) RC delay between this work and ref [3]. The optimized RC delay is 2 magnitudes smaller and falls in a few ps regime, while there are only minor differences in terms of switching speed for devices with comparable area. Therefore, the measured switching speed by NLS model is clarified as FE material property instead of being overwhelmed by device RC delay.

Fig. 11 presents the switching speed of crossbar array and single crossbar devices with comparable total area. It is clear that smaller size MFM with array structure boosts switching speed, while keeping the total area large enough for accurate current measurements. The switching speed in smaller array devices is faster while the parasitic effect is similar for both structures. This is because the switch speed is determined by domain wall propagation speed in FE HZO.

Conclusion

Sub-μm crossbar array FE HZO MFM devices were fabricated and the polarization switching of these devices is studied. Record fast polarization switch speed of 360 ps is obtained. The work unveils that domain wall propagation speed in HZO is the limiting factor for switch speed and more aggressively scaled devices will offer much faster switch speed. The work is supported by SRC JUMP ASCENT Center.

References

[1] J. Muller *et al.*, *Nano Lett.*, p.4318, 2012. [2] K. Ni *et al.*, *IEEE TED*, p. 2461, 2018. [3] X. Lyu *et al.*, *IEDM*, p. 342, 2019. [4] W. Chung *et al.*, *VLSI*, p. T89, 2018. [5] M. Si et al., *APL,* p. 072107, 2019. [6] E. Yurchuk *et al.*, *IEEE TED*, p. 3699, 2014. [7] J. Matthew *et al.*, *IEDM*, p. 139, 2017. [8] H. K. Yoo *et al.*, *IEDM*, p. 481, 2017. [9] S. Dunkel *et al.*, *IEDM*, p. 485, 2017. [10] C. Alessandri *et al.*, *IEEE EDL*, p. 1780, 2018. [11] K. Karda *et al.*, *IEEE EDL*, p. 801, 2016. [12] J. Y. Jo *et al.*, *PRL*, p. 267602, 2007.

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Campbell, Jason; Cheung, Kin (Charles); Ye, Peide. "Record Fast Polarization Switching Observed in Ferroelectric Hafnium Oxide Crossbar Arrays." Paper presented at IEEE Silicon Nanoelectronics Workshop, Honolulu, HI, US. June 13, 2020 - June 14, 2020.

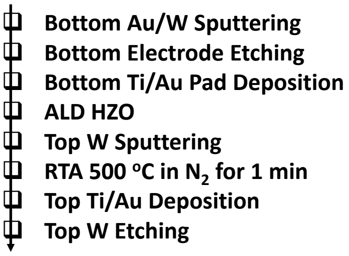


Fig. 1. Device fabrication process flow for W/HZO/W crossbar arrays and single crossbars.

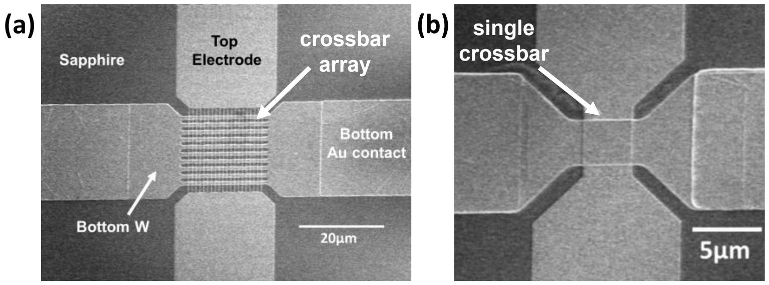


Fig. 2. SEM image of (a) a crossbar array device and (b) a single crossbar device. Insulating sapphire substrate is used for device fabrication considering on ultrafast measurements.

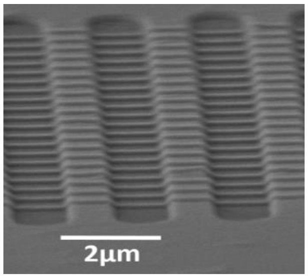


Fig. 3. SEM image of device structure details of a crossbar array device.

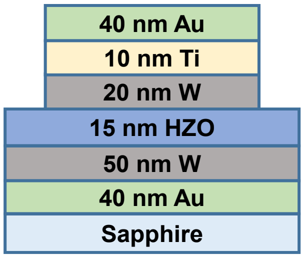


Fig.4. Cross sectional schematic diagram of W/HZO/W device structure.

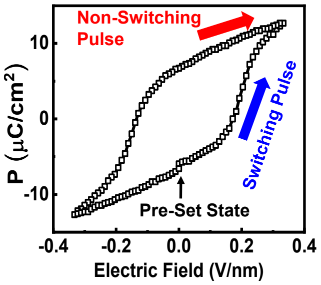


Fig. 5. P-E characteristics of a representative single crossbar MFM device with 15 nm thick HZO.

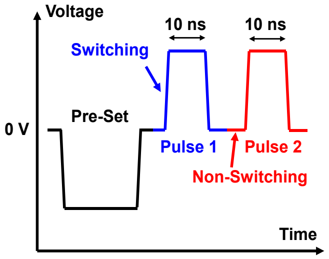


Fig. 6. PUND pulse sequence for transient ferroelectric polarization switching current measurement.

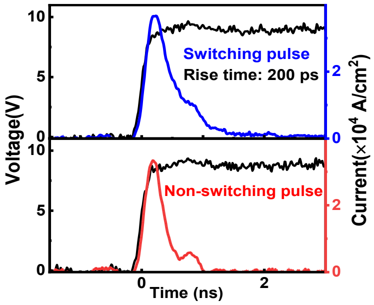


Fig. 7. Ultrafast pulse input with a 200 ps rise time and the corresponding transient current response of the switching (upper) and non-switching (lower) pulse.

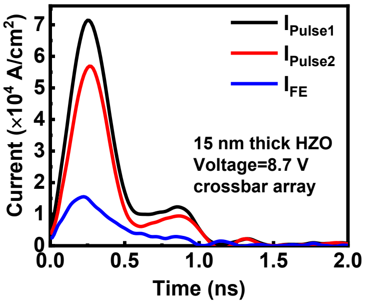


Fig. 8. Measured transient currents and the extracted switching current of a 15 nm thick HZO crossbar array device.

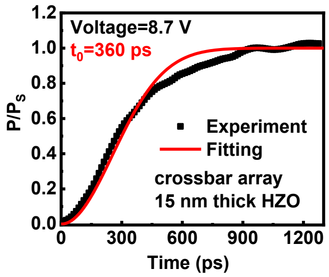


Fig. 9. Normalized transient switched polarization charge density from both experiment and fitting by NLS model of a 15 nm thick HZO crossbar array device.

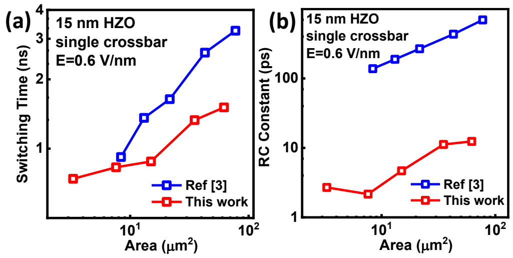


Fig. 10. Comparison of (a) switching time constant and (b) RC delay constant between the single crossbar devices with fabrication process improvement in this work and Ref. 3. The MFM area is similar.

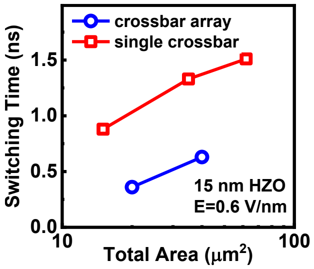


Fig. 11. Switching speed difference of single crossbar and crossbar array devices with similar total area.

Shrestha, Pragya; Lyu, xiao; Si, Mengwei; Campbell, Jason; Cheung, Kin (Charles); Ye, Peide. "Record Fast Polarization Switching Observed in Ferroelectric Hafnium Oxide Crossbar Arrays." Paper presented at IEEE Silicon Nanoelectronics Workshop, Honolulu, HI, US. June 13, 2020 - June 14, 2020.

Computational Process Control Compatible Dimensional Metrology Tool: Through-focus Scanning Optical Microscopy

Ravi Kiran Attota
Microsystem and Nanotechnology Division
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
Ravikiran.attota@nist.gov

Abstract—**Using only two derived numbers based on a reference library, this paper shows how through-focus scanning optical microscopy (TSOM) is compatible with computational process control (CPC) for the complete 3D shape process monitoring of nanoscale to microscale targets. This is demonstrated using three types of targets with widths (CDs) and depths ranging from 50 nm to 1.0 μm, and 70 nm to 20 μm, respectively. TSOM is a high-throughput, low-cost and in-line capable optical dimensional metrology method ideally suited for high-volume manufacturing (HVM), complementing other widely used metrology tools.**

Keywords—TSOM, process control, three-dimensional shape, optical microscope, through-focus, Computational Process Control

I. Introduction

As the applications of nanotechnology and micro-technology become widespread, usage of three-dimensional (3D) structures is increasing [1-10]. In addition, increased complexity and the resulting big data during manufacturing processes have resulted in a new paradigm: Computational Process Control (CPC) [11]. In the CPC environment, in-line metrology tools are highly desirable. Here, we present the through-focus scanning optical microscope (TSOM) as a dimensional metrology tool that is ideally compatible with CPC, and has many favorable attributes as a metrology and process control tool [12-21].

II. TSOM

TSOM is a method that collects and exploits the entire through-focus optical intensity information in 3-D space using a conventional optical microscope. Developments in image acquisition techniques have significantly reduced the acquisition time for a set of through-focus images so that it can be as fast as a single conventional microscope image, making TSOM suitable for HVM [22]. A vertical cross-section extracted from this 3D data results in a TSOM image. D-TSOM images are generated by taking a pixel-by-pixel difference between images of two separate targets. D-TSOM images can reveal sub-nanometer differences between nominally identical targets. The color (intensity) patterns of D-TSOM images are usually distinct for different types of parameter changes and serve as a “fingerprint” for different types of parameter variations, while remaining qualitatively similar for different magnitude changes in the same parameter. The magnitude of the optical content of D-TSOM images is proportional to the magnitude of the dimensional differences. The Optical Intensity Range (OIR, the difference between the maximum and the minimum optical intensity, multiplied by 100), provides a quantitative estimate of the difference between the two TSOM images. The utility of D-TSOM is that the color pattern of the D-TSOM image is an indicator of the difference in 3D shape, while the magnitude of the OIR scales the dimensional difference between the two targets.

III. Experiments

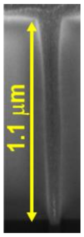


Fig.1

As a demonstration of the application of TSOM to high-aspect ratio (HAR) structures, in an SiO_2 layer on a 300 mm Si substrate with nominally 100 nm CD, 1100 nm depth, and 1000 nm pitch were studied, using a horizontal field-of-view (FOV) of 50 μm. A typical focused ion beam (FIB) cross-sectional view of the HAR target is shown Fig. 1. A mosaic of D-TSOM images obtained for the entire wafer (with central die as a reference target) is presented in Fig. 2(a). From this, four major types of D-TSOM image patterns (Fig. 2(b)) can be identified with their corresponding FIB cross-sectional profile differences as shown in Fig. 2(c).

IV. Process control results

If one considers the information available in Figs. 2(b) and 2(c) as a library, a simple process monitoring procedure can be proposed as shown in Figs. 3 and 4, based on the following rules selected for this demonstration. If the OIR of the D-TSOM image is more than 12, reject the target, as the dimensional differences are more than the tolerable limits. On the lower side, if the OIR of the D-TSOM image is less than 7, accept the target, as the dimensional differences are within the acceptable

Attota, Ravikiran. ”Computational Process Control Compatible Dimensional Metrology Tool: Through-focus Scanning Optical Microscopy.”
Paper presented at 2020 31st Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), Saratoga Springs, NY, US. August 24, 2020 - August 26, 2020.

level. If the OIR value is in between 7 and 12, accept the production target if the profiles are symmetric (T1 and T3) and reject if the profiles are asymmetric (T2 and T4). Selection of the rules depends on the desired outcome from the fabrication.

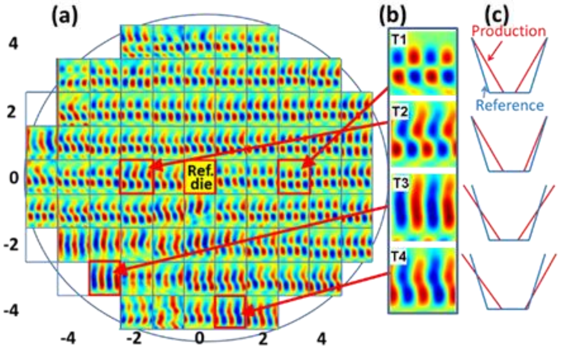
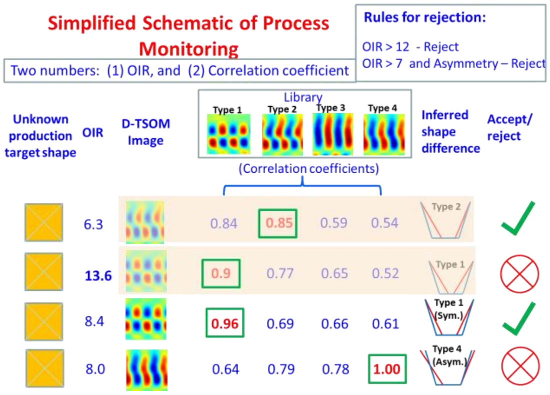


Fig. 2. (a) A mosaic of the D-TSOM images obtained by subtracting the TSOM image of the central reference target from the TSOM images of the targets in the other dies (with color scale bar set to automatic). (b) Four major types (T1, T2, T3 and T4) of D-TSOM image color patterns are identified. (c) *Schematic cross-sectional profile differences corresponding to (b) obtained from FIB cross-sectional analysis.* Nominal values: pitch = 1,000 nm, CD = 100 nm, and depth = 1100 nm. Illumination wavelength = 520 nm, numerical apertures (NA) = 0.75, illumination NA (INA) = 0.25.



Simplified Schematic of Process Monitoring

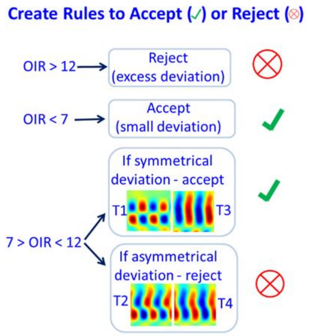
Two numbers: (1) OIR, and (2) Correlation coefficient

Rules for rejection:
OIR > 12 - Reject
OIR > 7 and Asymmetry – Reject

Unknown production target shape	OIR	D-TSOM Image	Type 1	Type 2	Type 3	Type 4	Inferred shape difference	Accept/ reject
	6.3		0.84	0.85	0.59	0.54	Type 2	✓
	13.6		0.9	0.77	0.65	0.52	Type 1	⊗
	8.4		0.96	0.69	0.66	0.61	Type 1 (Sym.)	✓
	8.0		0.64	0.79	0.78	1.00	Type 4 (Asym.)	⊗

Fig. 3. Proposed TSOM-based automated 3-D-shape process control method. The selected test production targets are considered to have unknown 3-D-shape profile (column 1). Comparing the correlation coefficients between the D-TSOM images of the test targets and the library provides the best match (green boxes) from which the possible 3-D shape difference type can be inferred (column 8).

Fig. 3 demonstrates that based on the type of 3-D shape difference, and the magnitude of the dimensional difference, the process control decision of accept/reject can be made based solely on the two numbers: OIR and correlation coefficient with minimal or no human intervention to make it suitable for CPC. A library and a customized rule set must be created beforehand. Note that the process control decision is automatically made on the very optical tool during measurement using the previously mentioned numbers. This method nearly eliminates the need for post-processing.



Create Rules to Accept (✓) or Reject (⊗)

OIR > 12 → Reject (excess deviation) ⊗

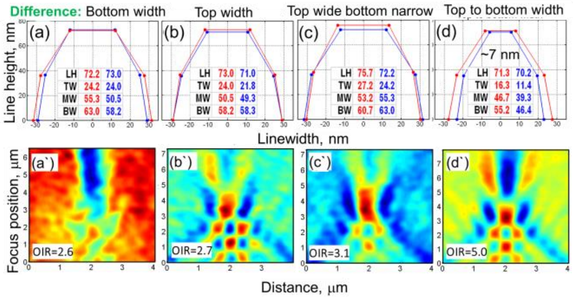
OIR < 7 → Accept (small deviation) ✓

7 > OIR < 12 → If symmetrical deviation - accept (T1, T3) ✓

7 > OIR < 12 → If asymmetrical deviation - reject (T2, T4) ⊗

Fig. 4. This schematic shows that only two numbers are needed to accomplish the 3D-shape process control with TSOM. OIR to determine the magnitude of the dimensional deviation, and correlation coefficient to determine the 3D shape similarity.

A second example is shown in Fig. 5 for substantially smaller, isolated line targets with nominal sizes of 70 nm height and 50 nm width [17]. In this case also, since TSOM can identify different types of 3D-shape differences (down to 0.8 nm), the similar procedure presented above can be applied for process control.



	(a)		(b)		(c)		(d)	
LH	72.2	73.0	73.0	71.0	75.7	72.2	71.3	70.2
TW	24.2	24.0	24.0	21.8	27.2	24.2	16.3	11.4
MW	55.3	50.5	50.5	49.3	53.2	55.3	46.7	39.3
BW	63.0	58.2	58.2	58.3	60.7	63.0	55.2	46.4

Fig. 5. Experimental 3D-shape analysis of isolated lines for process control. (a,b,c,d) CD-AFM-measured plots showing different types of 3D-shape differences. (a`,b`,c`,d`) are the corresponding D-TSOM images producing noticeably different color patterns. CD-AFM measured values of line height (LH), top width (TW), middle width (MW) and bottom width (BW, in nanometers) are shown in order for the pair of lines. Nominal height = 70 nm, and width = 50 nm.

A third example for a larger through-silicon via (TSV) target with a nominal diameter and depth of 1 µm and 20 µm respectively, is shown in Fig. 6. The similarity of the two D-TSOM images on the right indicates a similar pattern of dimensional differences. In this case also, the similar procedure presented above can be applied for process control using a library.

V. POTENTIAL APPLICATIONS

TSOM is a high-throughput, low-cost, nondestructive, robust, and easy-to-use 3D shape metrology method. It has 1 nm or better measurement resolution[14] for target sizes (depths/heights) ranging from sub-nanometer to over 100 um (over five orders of magnitude in size range). One of the unique characteristics of the TSOM method is its ability to reduce or eliminate optical cross correlations between different parameters, resulting in reduced measurement uncertainty. All the targets within the field-of-view can be analyzed simultaneously, including isolated structures with random shapes, repeated and non-repeated structures, and a variety of target materials. TSOM complements other tools by filling in some gaps in capabilities.

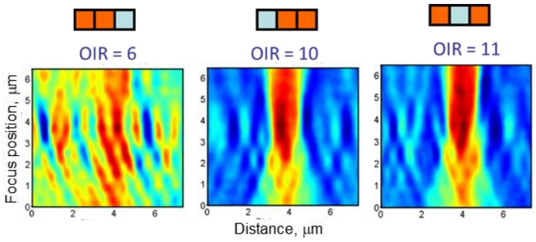


Fig. 6. Measured D-TSOM images representing the pairwise differences between TSOM images of three TSVs in three different dies. Nominal diameter = 1 µm, and depth = 20 µm.

TSOM has been successfully employed for many applications. Demonstrated applications of TSOM include critical dimension (linewidth), overlay (including over 10 um separation), patterned defect detection and analysis, FinFETs, nanoparticles (including soft-nanoparticles), photo-mask, thin-films (less than 0.5 nm to 10 nm) thickness, 2D materials, through-silicon vias (TSVs), high-aspect-ratio (HAR) targets, MEMS/NEMS devices, micro/nanofluidic channels, flexible electronics and micro-resonators/frequency combs. Other potential applications include three-dimensional shape process monitoring of self-assembled nanostructures, waveguides, and nanoimprint targets. Numerous industries benefit from the TSOM method —such as the semiconductor industry, flexible electronics, quantum science, photovoltaics, display, nanotechnology, MEMS, NEMS, biotechnology, data storage, and photonics.

VI. SUMMARY

Here we have briefly presented potential applications of TSOM using a reference library. TSOM could fill some gaps by satisfying the HVM metrology needs of 3-D structures for which 3-D shape process control/monitoring is needed, complementing other widely used metrology tools. We also show that TSOM as presented here has a high compatibility with computational process control.

ACKNOWLEDGEMENTS

The author would like to thank and acknowledge the contributions made by Hyeonggon Kang, Keana Scott, Richard Allen, Andras E. Vladar, Benjamin Bunday, Emil Agocs and John Kramar.

REFERENCES

1. M. H. Postek, Robert, "Instrumentation and metrology for nanotechnology," 2004).
2. G. B. Picotto, L. Koenders, and G. Wilkening, "Nanoscale metrology," Measurement Science and Technology **20**, 080101 (2009).
3. R. K. Leach, R. Boyd, T. Burke, H. U. Danzebrink, K. Dirscherl, T. Dziomba, M. Gee, L. Koenders, V. Morazzani, A. Pidduck, D. Roy, W. E. S. Unger, and A. Yacoot, "The European nanometrology landscape," Nanotechnology **22**(2011).
4. G. Häusler and S. Ettl, "Limitations of Optical 3D Sensors," in *Optical Measurement of Surface Topography*, R. Leach, ed. (Springer Berlin Heidelberg, 2011), pp. 23-48.
5. X. Zhang, H. Zhou, Z. Ge, A. Vaid, D. Konduparthi, C. Osorio, S. Ventola, R. Meir, O. Shoval, R. Kris, O. Adan, and M. Bar-Zvi, "Addressing FinFET metrology challenges in 1× node using tilt-beam critical dimension scanning electron microscope," MOEMS **13**, 041407-041407 (2014).
6. I. Schulmeyer, L. Lechner, A. Gu, R. Estrada, D. Stewart, L. Stern, S. McVey, B. Goetze, U. Mantz, and R. Jammy, "Advanced metrology and inspection solutions for a 3D world," in *2016 International Symposium on VLSI Technology, Systems and Application (VLSI-TSA)*, 2016), 1-2.
7. A. Arceo, B. Bunday, A. Cordes, and V. Vartanian, "Evolution or revolution: the path for metrology beyond the 22nm node," Solid State Technol **55**, 15-19 (2012).
8. T. F. Crimmins, "Defect metrology challenges at the 11-nm node and beyond," in 2010), 76380H-76380H-76312.
9. K. Takamasu, Y. Iwaki, S. Takahashi, H. Kawada, M. Ikota, G. F. Lorusso, and N. Horiguchi, "3D-profile measurement of advanced semiconductor features by reference metrology," in 2016), 97781T-97781T-97788.
10. R. Attota, R. Silver, and B. M. Barnes, "Optical through-focus technique that differentiates small changes in line width, line height and sidewall angle for CD, overlay, and defect metrology applications," Proc. SPIE **6922**, 69220E (2008).
11. D. Lammers, "COMPUTATIONAL PROCESS CONTROL SOLUTIONS TO SERVICE FABS IN CHINA" (2018), retrieved http://www.appliedmaterials.com/nanochip/nanochip-fab-solutions/march-2018/computational-process-control-solutions-to-service-fabs-in-china.
12. R. Attota, T. A. Germer, and R. M. Silver, "Through-focus scanning-optical-microscope imaging method for nanoscale dimensional analysis," Opt Lett **33**, 1990-1992 (2008).
13. R. Attota and R. Silver, "Nanometrology using a through-focus scanning optical microscopy method," Meas Sci Technol **22**(2011).
14. R. Attota, B. Bunday, and V. Vartanian, "Critical dimension metrology by through-focus scanning optical microscopy beyond the 22 nm node," Appl Phys Lett **102**(2013).
15. M. Ryabko, S. Koptyaev, A. Shcherbakov, A. Lantsov, and S. Y. Oh, "Motion-free all optical inspection system for nanoscale topology control," Opt Express **22**, 14958-14963 (2014).
16. S. Han, T. Yoshizawa, S. Zhang, J. H. Lee, J. H. Park, D. Jeong, E. J. Shin, and C. Park, "Tip/tilt-compensated through-focus scanning optical microscopy," Proc. of SPIE **10023**, 100230P (2016).
17. R. Attota and R. G. Dixson, "Resolving three-dimensional shape of sub-50 nm wide lines with nanometer-scale sensitivity using conventional optical microscopes," Appl Phys Lett **105,** 043101, (2014).
18. R. Attota, R. G. Dixson, J. A. Kramar, J. E. Potzick, A. E. Vladar, B. Bunday, E. Novak, and A. Rudack, "TSOM Method for Semiconductor Metrology," Proc. SPIE **7971**, 79710T (2011).
19. A. Arceo, B. Bunday, V. Vartanian, and R. Attota, "Patterned Defect & CD Metrology by TSOM Beyond the 22 nm Node," Proc Spie **8324**(2012).
20. S.-w. Park, G. Park, Y. Kim, J. H. Cho, J. Lee, and H. Kim, "Through-focus scanning optical microscopy with the Fourier modal method," Opt Express **26**, 11649-11657 (2018).
21. S. I. Association, "The International Technology Roadmap for Semiconductors (ITRS) " (Semiconductor Industry Association, San Jose, 2016).
22. R. Attota, "Through-focus or volumetric type of optical imaging methods: a review," J Biomed Opt **23**, 19100-19114 (2018).
23. R. K. Attota, H. Kang, K. Scott, R. Allen, A. E. Vladar, and B. Bunday, "Nondestructive shape process monitoring of three-dimensional, high-aspect-ratio targets using through-focus scanning optical microscopy," Measurement Science and Technology **29**, 125007 (2018).

Ray-based classification framework for high-dimensional data

Justyna P. Zwolak
National Institute of
Standards and Technology
Gaithersburg, MD 20899
jpzwolak@nist.com

Sandesh S. Kalantre
Joint Quantum Institute
University of Maryland
College Park, MD 20742
skalantr@umd.edu

Thomas McJunkin
Department of Physics
University of Wisconsin
Madison, WI 53706
tmcjunkin@wisc.edu

Brian J. Weber
Institute for Mathematical Sciences
ShanghaiTech University
Shanghai, 201210, China
bjweber@shanghaitech.edu.cn

Jacob M. Taylor
Joint Center for Quantum Information
and Computer Science
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
jacob.taylor@nist.gov

Abstract

While classification of arbitrary structures in high dimensions may require complete quantitative information, for simple geometrical structures, low-dimensional qualitative information about the boundaries defining the structures can suffice. Rather than using dense, multi-dimensional data, we propose a deep neural network (DNN) classification framework that utilizes a minimal collection of one-dimensional representations, called *rays*, to construct the "fingerprint" of the structure(s) based on substantially reduced information. We empirically study this framework using a synthetic dataset of double and triple quantum dot devices and apply it to the classification problem of identifying the device state. We show that the performance of the ray-based classifier is already on par with traditional 2D images for low dimensional systems, while significantly cutting down the data acquisition cost.

1 Introduction

Deep learning, with its remarkable progress in recent years [1, 2], is ripe for applications in physics [3]. A particular instance having general applicability to physical problems is the classification of arbitrary convex geometrical shapes embedded in an N-dimensional space [4]. Having a mathematical framework to understand this class of problems and a solution that scales efficiently with the dimension N is essential. With increasing effective dimensionality of the system, including parameters and data, determining the geometry with measurements across the full parameter space may become prohibitively expensive. However, as we show, qualitative information about the boundaries defining the structures of interest may suffice for classification.

We propose a new framework for classifying simple high-dimensional geometrical structures: *ray-based classification*. Rather than working with the full N-dimensional data tensor, we train a fully connected DNN using one-dimensional representations in $\mathbb{R}^N$, called "rays", to recognize the relative position of features defining a given structure. We position the boundaries of this structure relative to a point of interest, effectively "fingerprinting" its neighborhood in the $\mathbb{R}^N$ space. The ray-based classifier is motivated primarily by experiments, particularly those in which sparse data collection is impractical. Our approach not only reduces the amount of data that needs to be collected, but also can be implemented *in situ* and in an online learning setting, where data is acquired sequentially.

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.

Zwolak, Justyna; Taylor, Jacob; Kalantre, Sandesh; McJunkin, Thomas; Weber, Brian. "Ray-based classification framework for high-dimensional data." Paper presented at Thirty-fourth Conference on Neural Information Processing Systems, Vancouver, CA. December 05, 2020 - December 12, 2020.

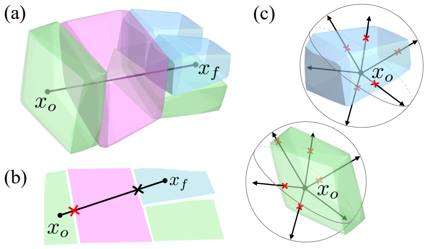


Figure 1: (a) Visualization of a ray $\mathfrak{R}(x_o, x_f)$ from x_o to x_f in $\mathbb{R}^3$. Different colors of polytopes represent different classes. (b) A side-view of the polytopes with two features marked along the ray. The red mark denotes a critical feature. (c) Visualization of the M-projection from point x_o with 6 rays (denoted by black arrows) for two different polytopes in $\mathbb{R}^3$. Note that both M-projections include a ray that does not have a critical feature.

We test the proposed framework using a modified version of the "Quantum dot data for machine learning" dataset [5] developed to study the application of convolutional neural networks (CNNs) to enhance calibration of semiconductor quantum dot devices for use as qubits [6]. Tuning these devices requires a series of measurements of a single response variable as a function of voltages on electrostatic gates. As the number of gates increases [7, 8], heuristic classification and tuning becomes increasingly difficult, as does the time it takes to fully explore the voltage space of all relevant gates. The specific geometry of the response in gate-voltage space corresponds to the number and position of populated quantum dots, which is valuable information in the process of tuning of these systems.

Previous work has shown both theoretically [9] and experimentally [10] that an image-based CNN classifier for 2D volumes, *i.e.*, solid images, combined with conventional optimization routines, can assist experimental efforts in tuning quantum dot devices between zero-, single- and double-dot states. Here, we consider a double- and triple-dot system. We show that using ray-based classification, the quantity of data required (and thus the time required) for identifying the state of the quantum dot system can be drastically reduced compared to an imaged-based classifier.

2 A Framework for Ray-Based Classification

Consider Euclidean space $\mathbb{R}^N$ with its conventional 2-norm distance function d, and a polytope function $p : \mathbb{R}^N \to \{0, 1\}$. The set of points where $p(x) = 1$ constitutes the boundary of a collection of polytopes. For example, a polytope function producing a square in $\mathbb{R}^2$ centered at the origin is $p(x_1, x_2) = \{1 \text{ if } |x_1| + |x_2| = 1; 0 \text{ elsewhere}\}$, where $(x_1, x_2) \in \mathbb{R}^2$. In our quantum dot applications a value of $p = 1$ indicates the location where an electron is transferred in or out of a dot.

Definition 1 (Rays). *Given $x_o, x_f \in \mathbb{R}^N$, the **ray** $\mathfrak{R}_{x_o,x_f}$ emanating from x_o and terminating at x_f is the set* $\{x \,|\, x = (1-t)x_o + tx_f, t \in [0, 1]\}$ (see Fig. 1(a) for a depiction of a ray in $\mathbb{R}^3$).

In practical applications, rays have a natural granularity that depends on the system as well as the data collection density. For quantum dots, the device parameters define an intrinsic separation between critical features that gives the scale of the problem. We refer to granularity of rays in terms of pixels.

To assess the geometry of a polytope enclosing any given point x_o, we consider a collection of rays of a fixed length r centered at x_o. The rays are uniquely determined by a set of M points on the sphere $\mathcal{S}_r^{N-1}$ of radius r centered at x_o, $\mathcal{P} := \{x_m \in \mathcal{S}_{x_o}^{N-1}(r) \,|\, 1 \leq m \leq M\}$. We call a set of M rays, $\mathcal{R}_M := \{\mathfrak{R}_{x_o,x_m} \,|\, x_m \in \mathcal{P}\}$, an ***M*-projection** (see Fig. 1(c) for visualization in $\mathbb{R}^3$).

Definition 2 (Feature). *Given a ray $\mathfrak{R}_{x_o,x_f}$ and a polytope function p, a point $x \in \mathfrak{R}_{x_o,x_f}$ is a **feature** if* $p(x) = 1$.

Figure 1(b) shows two features along a sample ray in $\mathbb{R}^3$. Features along a given ray define its ***feature set***, $F_{x_o,x_f} := \{x \in \mathfrak{R}_{x_o,x_f} \,|\, p(x) = 1\}$, with a natural order given by the 2-norm distance function $d : x_o \times F_{x_o,x_f} \to \mathbb{R}^+$. In general, F_{x_o,x_f} could be empty. Using a decreasing ***weight function*** $\gamma : \mathbb{R}^+ \to [0, 1]$ we can assign a weight to each feature, effectively defining the ***weight set*** Γ_{x_o,x_f} corresponding to its feature set F_{x_o,x_f} as $\Gamma_{x_o,x_f} = \{\gamma(d(x, x_o)) \,|\, x \in F_{x_o,x_f}\}$. The actual choice of function γ needs be altered to fit the problem itself and can be considered another hyperparameter that can help optimize the machine learning process. For the quantum dot case, we chose $\gamma(n) = {}^1\!/_n$.

The assumption that the weight function γ is monotonic in distance lets us define a ray's ***critical feature*** as the point $x \in F_{x_o,x_f}$ with highest (i.e., ***critical***) weight $W_{x_o,x_f} = \gamma(d(x, x_o))$. If $F_{x_o,x_f} = \varnothing$, we put $W_{x_o,x_f} = 0$. This allows us to "fingerprint" the space surrounding point x_o.

Definition 3 (Point fingerprint)**.** *Let $x_o \in \mathbb{R}^N$ be a point from which a collection of rays $\mathcal{R}_M = \{\mathfrak{R}_{x_o,x_f^1}, \ldots, \mathfrak{R}_{x_o,x_f^M}\}$ emanate. The* ***point fingerprint*** *of x_o is the M-dimensional vector consisting of the rays' critical weights: $\mathcal{F}_{x_o} = \left(W_{x_o,x_f^1}, \ldots, W_{x_o,x_f^M}\right)$.*

This point fingerprint $\mathcal{F}_{x_o}$ of x_o is the primary object of the ray-based classification framework. If sufficiently many rays in appropriate directions are chosen from x_o, the fingerprint is sufficient, at least in principle, to qualitatively determine the geometry of the convex polytope enclosing x_o. Due to the cost of experimental data acquisition, determining *how few* rays are sufficient for a machine learning algorithm to make this determination is of crucial importance. Looking to establish a correspondence between the fingerprint $\mathcal{F}_{x_o}$ of point x_o and the class of the polytope enclosing this point, we define the following problem:

Problem 1. *Given a set of bounded and unbounded convex polytopes filling an N-dimensional space and belonging to C distinct classes, $C \in \mathbb{N}$, and a point $x_o \in \mathbb{R}^N$, determine to which of the classes the polytope enclosing x_o belongs.*

Algorithm 1 Ray-based fingerprinting algorithm

```
Step 1. Find $M$-projection centered at $x_o$ given $r$.
Input: $x_o$, $r$, a set $\mathcal{P}$ of $M$ points on the $(N-1)$-sphere
$m \leftarrow 1$; $\mathcal{R}_M \leftarrow$ empty list
for $m = 1$ to $M$ do
    Find $m$-th ray $\mathfrak{R}_{x_o,x_f^m}$ and append it to the list $\mathcal{R}_M$.
end for
Return: List of $M$ rays $\mathcal{R}_M$.
Step 2. Fingerprint $x_o \in \mathbb{R}^N$ using rays in $\mathcal{R}_M$ from Step 1.
Input: $\mathcal{R}_M$, $\gamma : \mathbb{R}^+ \to [0,1]$
$m \leftarrow 1$; $\mathcal{F}_{x_o} \leftarrow$ empty list
for $m = 1$ to $M$ do
    Find the feature set $F_{x_o,x_f^m}$.
    if $F_{x_o,x_f^m} \neq \varnothing$ then
        Identify the critical feature $x_i^m$, find $W_{x_o,x_f^m}$ and
        append it to the list $\mathcal{F}_{x_o}$.
    else
        Append 0 to the list $\mathcal{F}_{x_o}$.
    end if
end for
Return: The point fingerprint vector $\mathcal{F}_{x_o}$.
```

A solution to this problem in the supervised learning setting can be obtained by training a DNN with the input being the point fingerprint and the output identifying an appropriate class. The procedural steps for the proposed classification algorithm for N-dimensional data in the form of pseudocode are presented in Algorithm 1.

3 Experiments: Classifying Shapes in 2D and 3D

The ray-based data is generated using a physics-based simulator of quantum dot devices [11]. An example of a simulated measurement, like the ones typically seen in the laboratory, is shown in

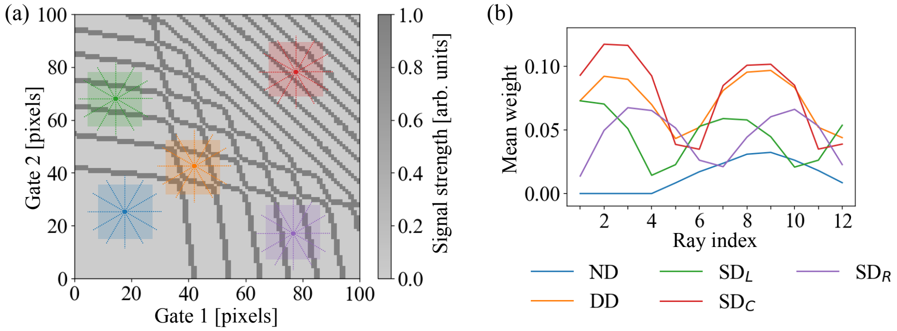


Figure 2: (a) A sample $2D$ map generated with the quantum dot simulator [11] showing the different bounded and unbounded polytopes in $\mathbb{R}^2$ with 12 evenly distributed rays overlaid on 2D scans like the ones used in Ref. [11]. (b) The average trends of the fingerprints with $M = 12$ rays. Fingerprints for SD$_L$ and SD$_R$ are out of phase, as expected from the the curvature of lines defining these states and SD$_C$ is shifted by $1/4$ of the period). The colors (labels) are consistent between both panels.

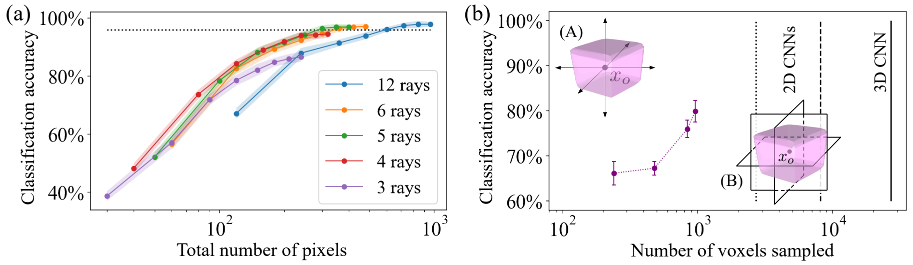


Figure 3: Classifier performance for varying numbers of rays as a function of the total number of (a) pixels measured and averaged over $N = 50$ training runs for the double-dot system and (b) voxel number averaged over $N = 10$ training runs for the triple-dot system. The black dashed line in (a) represents the benchmark from Ref. [11]. The black vertical lines in (b) represent the minimum data requirements for CNN classifier with 3 orthogonal 2D slices (as depicted in insert (B), dotted line), large 2D scan (dashed line), and a full 3D CNN (solid line). Insert (A) shows the M-projection with 6 rays. In both panels, the connecting lines are a guide to the eye only and the $3\,\sigma$ confidence bands.

Fig. 2(a). The x and y axes represent a subset of parameters that can be changed in the experiments (here, gate voltages) and the curves where the signal strength is equal to 1 represent the device response to a change in electron occupation. The slopes of those lines correspond to the location of the quantum dots with respect to the gates. The device states manifest themselves by different bounded and unbounded shapes defined by these curves, as shown in Fig. 2(a). Previous work has confirmed the reliability of a dataset generated with this simulator for the case of a CNN used with $2D$ images, finding an accuracy of $95.9\,\%$ (standard deviation $\sigma = 0.6\,\%$) over 200 training and validation runs performed on distinct datasets [11]. Here, we use a modified version of this dataset, splitting the single-dot (SD) class into 3 distinct classes based on the dot location (Left, Center, Right) as suggested by experimentalists. No-dot (ND) and double-dot (DD) classes are unchanged.

To test the ray-based classification framework in 2D, we use 20 realizations of $2D$ maps qualitatively comparable to the one shown in Fig. 2(a). Using a synthetic dataset allows us to systematically vary the length of the rays and their number. A regular grid of 1,369 points is used for sampling, resulting in a dataset of 27,380 fingerprints. We consider five datasets of M-projections, with $M = 3, 4, 5, 6, \mathrm{and} 12$ evenly spaced rays. The ray length is varied between 10 and 80 pixels (where 30 pixels is the average separation between transition lines in the simulated devices). We ran 50 training and validation tests per combination of rays' number and length (with data divided 80:20). For testing, we generated a separate dataset based on three distinct devices. This allows us to both better determine the classification error for the most efficient number and length combinations of rays and to study the failure cases over the device layout (see Fig. A.1).

Figure 3(a) shows the performance of the ray-based classifier. The accuracy of the classifier increases with the total number of points measured for a fixed number or rays, as expected. However, for a fixed number of points, increasing the number of rays does not necessarily lead to increased accuracy. This is because with a fixed number of points and point density, increasing the number of rays naturally results in shorter rays. Rays shorter than the radius of the interior diameter of the shapes leads to empty feature sets, resulting in uninformative fingerprints. Increasing the number or size of hidden layers in the DNN does not further improve the accuracy (see Table A.1).

To test the proposed framework with triple-dot systems [12], we generated a dataset by sampling 17,576 fingerprints from a single simulated device with three dot gates. We varied the number of rays between 6 and 18, while keeping the length of the rays fixed at 60 voxels. For each configuration, we performed $N = 10$ training and validation runs (with data divided 80:20). As shown in Fig. 3, the classifier accuracy improved from 66.2 % (σ = 0.3 %) for 6 rays to 79.9 % (σ = 0.3 %) for 18 rays.

4 Summary

While our analysis is performed using simulated data, its true advantage becomes clear when put in the context of experiments. Depending on the resolution, measuring 12 rays of length 80 (total of

960 data points) is equivalent to measuring the full $2D$ image (900 data points) as in Ref. [10]. With 6 rays of length 60 pixels, only 360 data points are needed, resulting in $60\,\%$ reduction of the data needed to obtain accuracy of $96.4\,\%$ ($\sigma = 0.4\,\%$). The reduction for 3D data is even more significant.

In conclusion, we have defined a framework to generate a low-dimensional representation of geometrical shapes in a high-dimensional space. We have empirically shown that the ray-based framework is an effective solution for cutting down the measurement cost while preserving high-accuracy of classification on the quantum dot dataset. If the ray-based classifier were implemented in a scheme to tune the double dot, as in Ref. [10], this reduction in data collection significantly improves its viability as a replacement to a hand-tuning scheme by a human operator. We expect this approach will find many related applications outside of the quantum dot domain.

Broader Impact

The authors believe that research presented in this paper does not have ethical aspects. This work furthers the efforts to automate and scale up quantum dot-based quantum computing to larger and more impactful devices. Our approach can also be extended to setups involving the estimation of quantum states in solid-state and atomic experiments, as well as tuning and scalability of other quantum computing architectures. A ray-based approach may find use classifying point clouds if sufficiently ordered. Extending our technique to include intersection points beyond the first critical feature may allow for identification of non-convex shapes in higher dimensions.

Acknowledgments and Disclosure of Funding

This research was sponsored in part by the Army Research Office (ARO), through Grant No. W911NF-17-1-0274. S.K. gratefully acknowledges support from the Joint Quantum Institute (JQI) – Joint Center for Quantum Information and Computer Science (QuICS) Lanczos graduate fellowship. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ARO, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright noted herein. Any mention of commercial products is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc. (2012).

[2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press (2016).

[3] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová. Machine learning and the physical sciences. *Rev. Mod. Phys.* **91**, 045002 (2019).

[4] P. Rotondo, M. C. Lagomarsino, and M. Gherardi. Counting the learnable functions of geometrically structured data. *Phys. Rev. Res.* **2**, 023169 (2020).

[5] Quantum dot data for machine learning, 2018. `https://catalog.data.gov/dataset/quantum-dot-data-for-machine-learning`.

[6] D. Loss and D. P. DiVincenzo. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126, (1998).

[7] D. M. Zajac, T. M. Hazard, X. Mi, E. Nielsen, and J. R. Petta. Scalable Gate Architecture for a One-Dimensional Array of Semiconductor Spin Qubits. *Phys. Rev. Appl.* **6**, 054013 (2016).

[8] U. Mukhopadhyay, J. P. Dehollain, Ch. Reichl, W. Wegscheider, and L. M. K. Vandersypen. A 2×2 quantum dot array with controllable inter-dot tunnel couplings. *Appl. Phys. Lett.* **112**, 183505 (2018).

Zwolak, Justyna; Taylor, Jacob; Kalantre, Sandesh; McJunkin, Thomas; Weber, Brian. "Ray-based classification framework for high-dimensional data." Paper presented at Thirty-fourth Conference on Neural Information Processing Systems, Vancouver, CA. December 05, 2020 - December 12, 2020.

[9] S. S. Kalantre, J. P. Zwolak, S. Ragole, X. Wu, N. M. Zimmerman, M. D. Stewart, and J. M. Taylor. Machine learning techniques for state recognition and auto-tuning in quantum dots. *npj Quantum Inf.* **5**, 1–10 (2019).

[10] J. P. Zwolak, T. McJunkin, S. S. Kalantre, J.P. Dodson, E.R. MacQuarrie, D.E. Savage, M.G. Lagally, S.N. Coppersmith, M. A. Eriksson, and J. M. Taylor. Autotuning of double-dot devices in situ with machine learning. *Phys. Rev. Appl.* **13**, 034075 (2020).

[11] J. P. Zwolak, S. S. Kalantre, X. Wu, S. Ragole, and J. M. Taylor. QFlow lite dataset: A machine-learning approach to the charge states in quantum dot experiments. *PLoS ONE*, **13**, 1–17 (2018).

[12] D. Schroer, A. D. Greentree, L. Gaudreau, K. Eberl, L. C. L. Hollenberg, J. P. Kotthaus, and S. Ludwig. Electrostatically defined serial triple quantum dot charged with few electrons. *Phys. Rev. B***76** 075306 (2007)

Appendix

Overview of failure modes

To better understand the failure cases of the ray-based classifier for the best rays' number and length combinations, we use three test datasets comprised of a regular grid of 1,369 points sampled over three devices distinct from those used for training and validation. Figure A.1 visualizes the classification success on a stability diagrams like the one in Fig. 2(a). The test devices are shown as rows and the different rays' numbers and lengths combinations are shown as columns. As can be seen in columns two and four, with $r = 50$ pixels the classifiers fails when a sampled point falls on the boundary lines for the polygons.

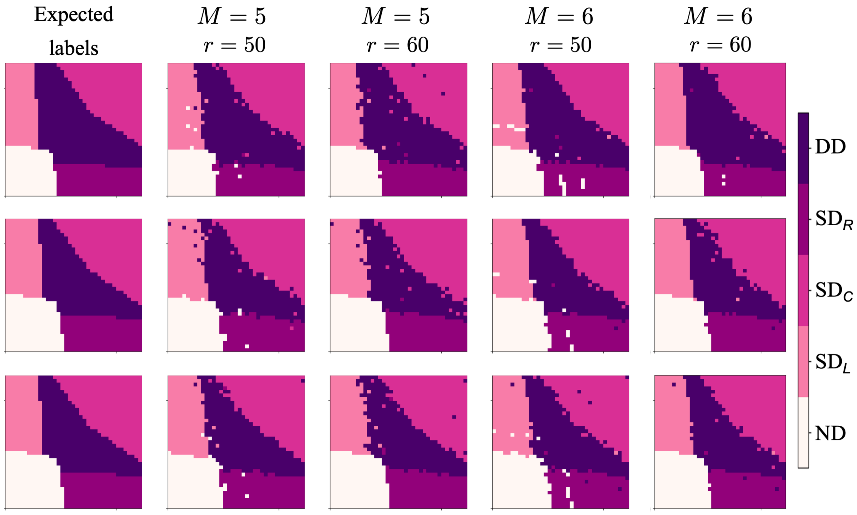


Figure A.1: Visualization of the failing cases overlaid on the test devices, with each row corresponding to a different device. The expected labels are shown in the leftmost column. The remaining columns are the ray-based classifier results for varying number of rays, M, and length of rays in pixels, r.

Analysis of DNN architectures

In experiments presented in Sec. 3 we use a DNN with 256, 128, and 32 neurons and the ReLU activation function for the fully connected hidden layers. The output layer consists of 5 neurons and the softmax activation function. We use the Adam optimizer ($\eta = 10^{-3}$), and the sparse categorical cross-entropy loss function. We found that increasing the size or complexity of the DNN (in terms of more or bigger layers) does not improve the performance of the ray-based classifier. In fact, for certain rays' number and length combination (e.g., 6 rays of length 60 pixels), a smaller network would suffice to achieve the same accuracy (see Table A.1).

Table A.1: Comparison of the varying DNN architectures for a fixed number and length of rays. For each network we report average accuracy, μ_r (%), and standard deviation, σ_r (%), where r denotes the length of the ray (in pixels), for $N = 50$ iterations of training and testing.

DNN	5 rays		6 rays	
	μ_{50} (σ_{50})	μ_{60} (σ_{60})	μ_{50} (σ_{50})	μ_{60} (σ_{60})
64-32	93.6 (0.4)	95.8 (0.4)	94.5 (0.3)	96.3 (0.3)
128-64-32	94.2 (0.4)	96.4 (0.4)	94.6 (0.4)	96.4 (0.4)
256-64-32	94.2 (0.5)	96.5 (0.4)	94.7 (0.4)	96.6 (0.3)
512-256-64-32	94.6 (0.4)	96.5 (0.4)	94.5 (0.4)	96.3 (0.3)

Operando Scanning Electron and Microwave Microscopies in Plasmas: A Comparative Analysis

Andrei Kolmakov[1*] *and Alexander Tselev*[2]

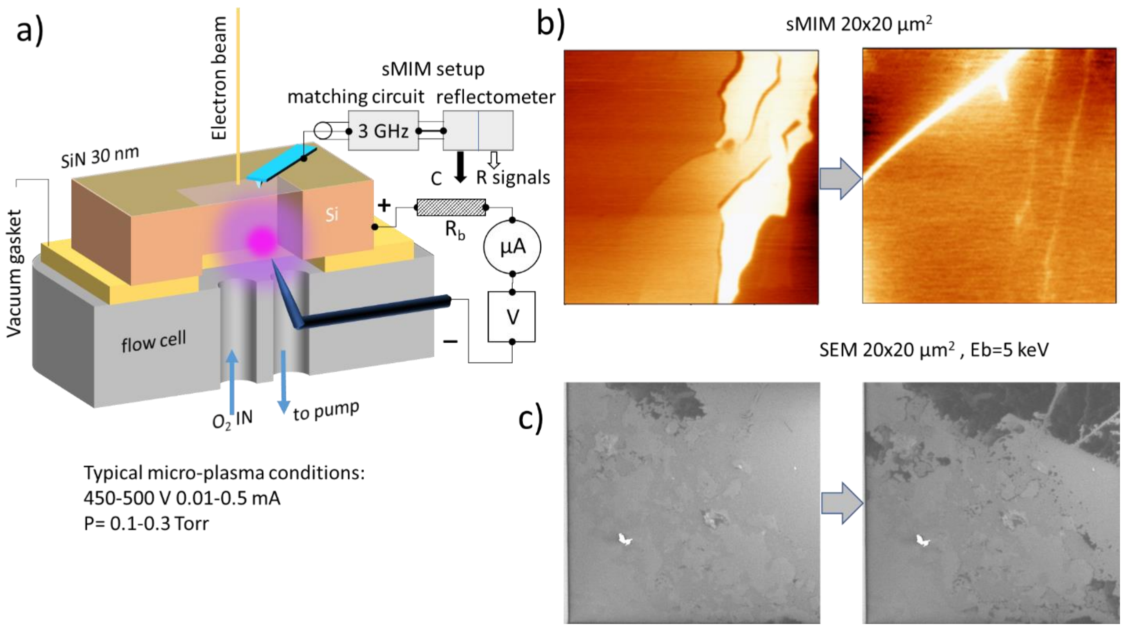
[1] Physical Measurements Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

[2] Department of Physics & CICECO–Aveiro Institute of Materials, University of Aveiro, 3810-193 Aveiro, Portugal

There exists a great need for an operando nanoscale characterization of evolution of surface composition and morphology during plasma-assisted processing. This includes sputter deposition, plasma-assisted etching, plasma-enhanced atomic layer deposition and other processes relevant for semiconductor and aerospace industries, environmental remediation, and biomedical technology. Recently, we proposed near-field scanning-probe-based microwave imaging as a tool to image surfaces immediately (a few seconds) after plasma processing with a sub-100 nm spatial resolution [1].

In this communication, we report on true operando near-field microwave imaging in plasma and its extension to imaging in SEM. The core of our approach is a microflow DC micro-discharge chamber equipped with a few-10s nm-thick SiN membrane transparent to a few keV electrons and microwave radiation (Figure 1a). This membrane isolates the probe of a scanning Microwave Impedance Microscope (sMIM) or an SEM column from plasma environment and thus enables real-time imaging of a surface of interest under plasma conditions. Using model systems, such as graphene, PMMA films, and polystyrene microparticles, we have comparatively explored performance of the SEM and sMIM in the same plasma-assisted processes. In particular, sensitivity, frame rate, spatial resolution, probing depth, and probe-induced effects were evaluated and compared. Figure 1b shows the process of etching of folded graphene by air DC plasma. The same process is depicted in the SEM snapshots in Figure 1c. The sMIM contrast in Figure 1b is due to graphene conductivity which degrades significantly after a few seconds of ca. 200 mW plasma treatment. Under optimal conditions, a resolution of ca. 50 nm can be achieved for metallic objects with commercially available probes. In turn, the contrast formation in SEM is due to local variations in the electron emission yield of the sample. This can be affected by presence of static charges in the SiN membrane. The spatial resolution can be as high as 10 nm with a frame rate on the order of 1 Hz. Interestingly, both the microscopies revealed formation of highly conductive filamentary structures of graphene during plasma etching—a phenomenon that is a subject of ongoing research.

Kolmakov, Andrei; Tselev, Alexander. "Operando Scanning Electron and Microwave Microscopies in Plasmas: A Comparative Analysis." Paper presented at Microscopy and Microanalysis 2020, Milwaukee, WI, US. August 02, 2020 - August 06, 2020.



Typical micro-plasma conditions:
450-500 V 0.01-0.5 mA
P= 0.1-0.3 Torr

Figure 1 (a) Experimental setup for operando sMIM and SEM studies under DC plasma; (b) sMIM (conductance channel) 20x20 μm^2 image of a graphene sample before (left panel) and after a few seconds (right panel) of application of DC air plasma (0.25 Torr, 500 V, 0.4 mA). The overall signal intensity drops (color scales are different in the images) due to local graphene etching, and filamentary structures are formed; (c) A similar sequence of 20x20 μm^2 images in SEM (5 keV) taken during DC (0.25 Torr, 450 V 0.2 mA) plasma treatment

References

[1] A. Tselev, J. Fagan, and A. Kolmakov, "In-situ near-field probe microscopy of plasma processing," *Applied Physics Letters,* vol. 113, no. 26, p. 263101, 2018.

GRADIENT ELUTION MOVING BOUNDARY ELECTROPHORESIS OF HOMEMADE FUEL-OXIDIZER EXPLOSIVES

Shannon T. Krauss[1], Dillon Jobes[2], and Thomas P. Forbes[1*]
[1]National Institute of Standards and Technology, USA
[2]Tulane University, USA

ABSTRACT

The analysis of complex samples without extensive preparation steps (*e.g.,* filtration or centrifugation) is a persistent challenge for applications ranging from forensic identification to high-throughput security screening. Here, a capillary-based microfluidic electrophoresis platform incorporating a hydrodynamic counterflow and capacitively coupled contactless conductivity detection is employed for the detection of inorganic components from propellants, pyrotechnics, and homemade fuel-oxidizer explosives. The bulk counterflow prohibits capillary clogging by particulate or fiber debris, as well as system fouling from fuels, *e.g.,* petroleum jelly or kerosene.

KEYWORDS: Capillary Electrophoresis, Gradient Elution Moving Boundary Electrophoresis, Capacitively Coupled Contactless Conductivity Detection, Propellants, Pyrotechnics, Fuel-Oxidizer Mixtures, Explosives

INTRODUCTION

Capillary electrophoresis (CE) is a selective analytical technique that has found extensive use for separations in biological applications [1]. As CE-based techniques have evolved, these systems have also found utility in applications ranging from separation of narcotic samples to analysis of post-blast debris. An array of configurations, modalities, and platforms have sought to take advantage of the benefits afforded by microfluidics, *e.g.,* automation, reduced sample sizes, multiplexing capabilities, and high-throughput analysis. One such method, gradient elution moving boundary electrophoresis (GEMBE), is a simple technique consisting of a relatively short capillary (5 cm) connecting two reservoirs and the application of an electric field (Figure 1(a)) [2]. Microfluidic electrophoresis platforms employing GEMBE separation modality have demonstrated the analysis of analytes ranging from amino acids to inorganics to opioids [2-4].

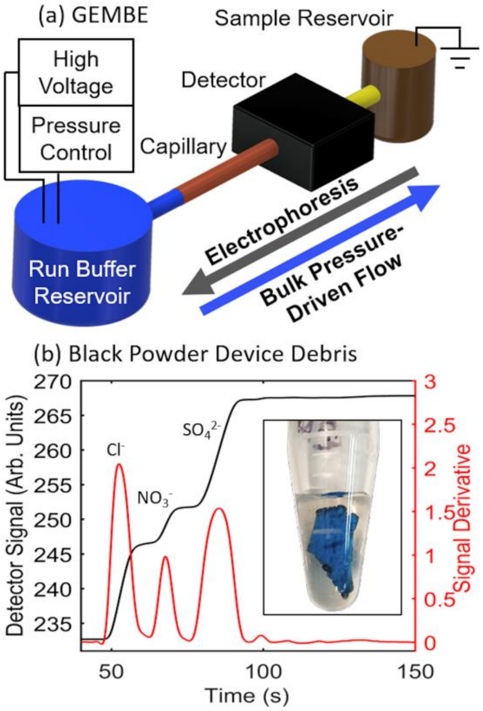


Figure 1: (a) Schematic representation of GEMBE separation. (b) GEMBE analysis of post-blast debris collected from a black powder charge thermoplastic polymer device. Raw signal showing step-wise increases in conductivity (black trace) and raw signal derivative demonstrating traditional peaks (red trace) from signal piece of debris (inset).

EXPERIMENTAL

GEMBE separation is achieved by gradually reducing a pressure-driven hydrodynamic counterflow (Figure 1(a)). As the hydrodynamic counterflow decreases, each analyte sequentially enters the capillary under electrophoretic transport. Each analyte is detected by coupled contactless conductivity detection (C^4D) as a moving boundary or step in conductivity (Figure 1(b)).

A 100 mmol/L acetic acid, 10 mmol/L L-histidine, pH 3.69 buffer and 200 µmol/L calcium chloride internal standard were used for separation under a -600 V/cm (-3000 V voltage) electric field. Following an elevated applied pressure (+30 kPa), the pressure was reduced at -100 Pa/s until all analytes were detected.

RESULTS AND DISCUSSION

The capabilities enabled by GEMBE provide unique utility for the direct analysis of inorganic components from homemade and improvised explosives and post-blast debris. GEMBE has the potential to separate and detect common inorganic oxidizers,

Forbes, Thomas P.; Krauss, Shannon; Jobes, Dillion. "Gradient Elution Moving Boundary Electrophoresis of Homemade Fuel-Oxidizer Explosives." Paper presented at MicroTAS 2020. October 04, 2020 - October 09, 2020.

including nitrate, chlorate, and perchlorate from low explosives without the need for extensive sample preparation or filtration. Figure 1(b) demonstrates the direct analysis of post-blast debris from the controlled heated wire detonation of a black powder (KNO_3/S/C) charge thermoplastic pipe-based improvised explosives device. Individual millimeter sized fragments of the device were analyzed without sample preparation, filtering, or centrifugation to eliminate particulate or other contaminants.

GEMBE separation of pre-blast black powders and black powder substitutes revealed detection of the main nitrate and perchlorate oxidizers without interference from an array of additional organic and inorganic fuels common to these propellants (*e.g.*, sulfur, dicyandiamide, nitro-benzoic acid, or sodium benzoate). Screening of tertiary explosives based on ammonium nitrate or potassium chlorate oxidizers with fuels including aluminum, icing sugar, petroleum jelly, fuel oil, and kerosene was also investigated. The GEMBE configuration excluded these fuels from fouling the capillary, enabled semi-quantitative analysis, and detection down to under 10 µmol/L.

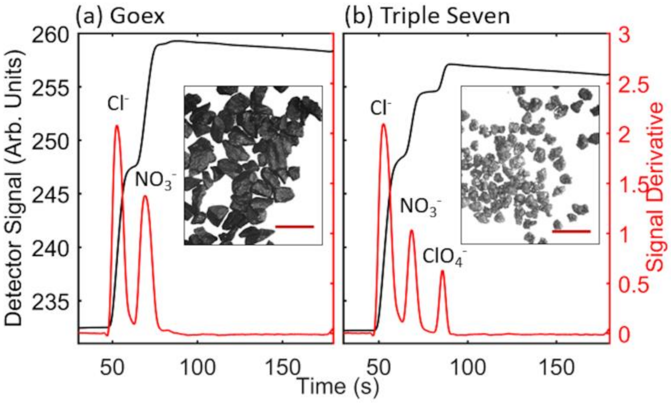


Figure 2: GEMBE separation of (a) Goex black powder and (b) Triple Seven black powder substitute samples containing nitrate and perchlorate. Raw data in black showing stepwise increases in conductivity and derivative representation of the raw data with identified peaks shown in red (internal standard: chloride). Bright field microscopy inset images of each with scale bar:2 mm.

CONCLUSION

GEMBE has the potential for quantitative forensic analysis and high throughput screening applications targeting the inorganic components of homemade explosives and post-blast debris. The nature of the GEMBE counterflow excludes particles, fibers, and detrimental components of samples from entering the capillary. This significantly reduces sample preparation (e.g., filtration or centrifugation) and capillary fouling. In addition, unlike traditional electrophoretic separations, GEMBE does not require a defined injection. This simplification allows for multiplexed devices, scaling up throughput. Furthermore, separation resolution is simply controlled through the rate of pressure reduction and applied voltage.

ACKNOWLEDGEMENTS

This research was performed while S.T.K. held a National Institute of Standards and Technology (NIST) National Research Council (NRC) Research Postdoctoral Associateship Award. The U.S. Department of Homeland Security Science and Technology Directorate sponsored a portion of the production of this material under Interagency Agreement IAA FTEN-18-00014 with the National Institute of Standards and Technology.

Certain commercial equipment, instruments, or materials are identified in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

REFERENCES

[1] M. Pioch, S.-C. Bunz, and C. Neusüß, "Capillary electrophoresis/mass spectrometry relevant to pharmaceutical and biotechnological applications," *Electrophoresis*, 33, 1517-1530, 2012.

[2] J. G. Shackman, M. S. Munson, and D. Ross, "Gradient Elution Moving Boundary Electrophoresis for High-Throughput Multiplexed Microfluidic Devices," *Anal. Chem.* 79, 565-571. 2007.

[3] E. A. Strychalski, A. C. Henry, and D. Ross, "Microfluidic Analysis of Complex Samples with Minimal Sample Preparation Using Gradient Elution Moving Boundary Electrophoresis," *Anal. Chem.* 81, 10201-10207, 2009.

[4] S. T. Krauss, D. Ross, and T. P. Forbes, "Separation and Detection of Trace Fentanyl from Complex Mixtures Using Gradient Elution Moving Boundary Electrophoresis," *Anal. Chem.* 91, 13014-13021, 2019.

CONTACT

* T.P. Forbes; phone: +1-301-975-2111; thomas.forbes@nist.gov

Forbes, Thomas P.; Krauss, Shannon; Jobes, Dillion. "Gradient Elution Moving Boundary Electrophoresis of Homemade Fuel-Oxidizer Explosives." Paper presented at MicroTAS 2020. October 04, 2020 - October 09, 2020.

NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware

Junxiao Shi
National Institute of Standards and Technology
Gaithersburg, MD, USA
junxiao.shi@nist.gov

Davide Pesavento
National Institute of Standards and Technology
Gaithersburg, MD, USA
davide.pesavento@nist.gov

Lotfi Benmohamed
National Institute of Standards and Technology
Gaithersburg, MD, USA
lotfi.benmohamed@nist.gov

ABSTRACT

Since the Named Data Networking (NDN) data plane requires name-based lookup of potentially large tables using variable-length hierarchical names as well as per-packet state updates, achieving high-speed NDN forwarding remains a challenge. In order to address this gap, we developed a high-performance NDN router capable of reaching forwarding rates higher than 100 Gbps while running on commodity hardware. In this paper we present our design and discuss its tradeoffs. We achieved this performance through several optimization techniques that include adopting better algorithms and efficient data structures, as well as making use of the parallelism offered by modern multi-core CPUs and multiple hardware queues with user-space drivers for kernel bypass. Our open-source forwarder is the first software implementation of NDN to exceed 100 Gbps throughput while supporting the full protocol semantics. We also present the results of extensive benchmarking carried out to assess a number of performance dimensions and to diagnose the current bottlenecks in the packet processing pipeline for future scalability enhancements. Finally, we identify future work which includes hardware-assisted ingress traffic dispatching, dynamic load balancing across forwarding threads, and novel caching solutions to accommodate on-disk content stores.

CCS CONCEPTS

• **Networks → Routers**; **Network performance analysis**; *Network layer protocols*; *Point-to-point networks.*

KEYWORDS

Named data networking, Information centric networking, NDN forwarder, Software router, Packet forwarding engine, High-speed forwarding, Network performance, Kernel bypass, Commodity hardware, Performance benchmarking

ACM Reference Format:
Junxiao Shi, Davide Pesavento, and Lotfi Benmohamed. 2020. NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware. In *7th ACM Conference on Information-Centric Networking (ICN '20), September 29–October 1, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3405656.3418715

ICN '20, September 29–October 1, 2020, Virtual Event, Canada
2020. ACM ISBN 978-1-4503-8040-9/20/09.
https://doi.org/10.1145/3405656.3418715

1 INTRODUCTION

Named Data Networking (NDN) is a new networking protocol with a data-centric communication architecture based on retrieval of named content rather than packet delivery between hosts [9, 39]. It is one of the most prominent instances of Information Centric Networking (ICN). Communication in NDN is receiver-driven: a consumer sends an Interest packet carrying the desired content name, the network uses this name to forward the request toward a producer or an in-network cache, and eventually a Data packet is returned to the consumer on the reverse path. A fundamental component in a NDN network is the *forwarder* (or router) that implements NDN's communication model following the behavior described by Shi [29, Chapter 3]. Accordingly, an NDN router must perform name-based lookups of potentially large tables using variable-length hierarchical names and simultaneously update its internal state on a per-packet basis. This makes wire-speed NDN forwarding challenging to achieve. At the same time, a number of scientific and other data-intensive applications [1, 4, 23, 27, 28] are hampered by the lack of such high-speed capability.

In this paper we present the design of **NDN-DPDK**, a high-performance NDN forwarder capable of achieving a throughput of over 100 Gbps while running on commodity x86 hardware. NDN-DPDK adopts several architectural optimizations ranging from better algorithms and data structures to reduced kernel and system call overhead, which was made possible by leveraging the fast user-space packet processing framework Data Plane Development Kit (DPDK) [17]. DPDK is available for many common 10/100 Gbps Ethernet adapters and provides a set of libraries to accelerate packet processing tasks, such as ring buffers, memory pools, and thread management. This enables our forwarder to receive and transmit packets directly from user space without going through the Linux kernel. Additionally, NDN-DPDK takes full advantage of the parallelism offered by modern multi-core CPUs.

Our open-source codebase, available on GitHub[1], possesses several unique features that advance the state of the art in NDN software routers:

- First, to the best of our knowledge, NDN-DPDK is the first complete implementation of a high-speed NDN forwarding engine on real hardware. Previous attempts [10, 11, 16, 22, 30, 33, 36] either focused on a subset of the data plane functions, did not support the full NDN protocol and name matching semantics, prioritized modularity and flexibility over performance, or relied on simulations rather than actual implementation.

[1] https://github.com/usnistgov/ndn-dpdk

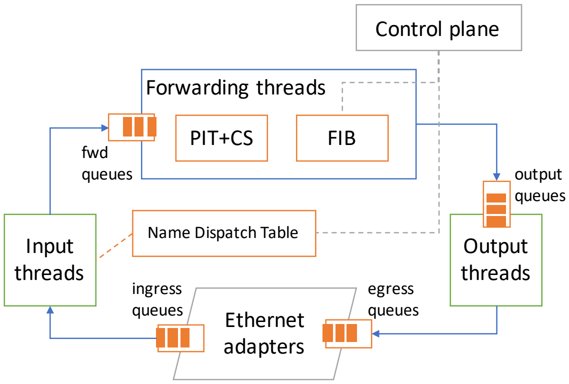


Figure 1: High-level view of NDN-DPDK's architecture.

- Second, since NDN allows for the Interest name to be a prefix of the Data name, we propose a novel approach for fast prefix matching between Interests and Data. An important ingredient of this approach is the PIT token, a small hop-by-hop header field added to each packet to accelerate PIT lookups.
- Third, in order to efficiently support prefix matching in NDN's Content Store, we developed a new solution based on indirect entries. Without such solution, an Interest carrying a non-exact name would have to be answered by the producer, reducing the effectiveness of in-network caching.

The rest of this paper is organized as follows. Section 2 provides an overview of NDN-DPDK's design. The forwarder's input and output stages are described in section 3, where packet dispatching for Interest and Data is also covered. Sections 4 and 5 introduce the main data structures and the details of their implementation. Support for NDN's forwarding strategies is discussed in section 6. The results of extensive performance benchmarks are presented in section 7. Finally, we review previous publications related to ICN forwarding in section 8, and conclude the paper in section 9, where we also list our future work.

2 DESIGN OVERVIEW

The forwarding plane of NDN-DPDK adopts a multi-threaded architecture (fig. 1). Each packet is processed in three stages:

1. The *input stage* receives a packet from a network interface, decodes it, and dispatches it to a forwarding thread according to a Name Dispatch Table (NDT).
2. The *forwarding stage* applies the NDN forwarding rules to the packet; this stage includes the traditional FIB (Forwarding Information Base), PIT (Pending Interest Table), and CS (Content Store) components, as well as the forwarding strategies.
3. The *output stage* prepares outgoing packets and passes them to a network interface for transmission.

This architecture allows the forwarder to make use of all available CPU cores and process several packets in parallel.

2.1 Memory Pools and NUMA Sockets

NDN-DPDK uses DPDK's *mempool* library to preallocate most of its data structures in 1 GiB hugepages. This eliminates the unpredictable latency of calling `malloc()` on the packet processing hot path, and reduces the complexity of handling runtime memory allocation failures. These memory pools are also pinned to physical memory pages and cannot be swapped out by the kernel, which ensures consistent access latency.

Most server-grade machines adopt a Non-Uniform Memory Access (NUMA) design. Under NUMA, each CPU, memory DIMM, and PCIe peripheral belongs to one *NUMA socket*. A CPU can access its local memory (memory located on the same NUMA socket) faster than non-local memory (memory located on a different NUMA socket). Using DPDK's *Environment Abstraction Layer* (EAL) library, NDN-DPDK pins its threads to specific CPU cores, and allocates most data structures used by a thread in NUMA-local memory. This minimizes memory access latency for these data structures.

Similarly, a PCIe Ethernet adapter also belongs to a NUMA socket. All packets received on a given adapter are stored in a memory pool local to that adapter's NUMA socket. Thus, NDN-DPDK assigns the input and output threads serving each interface to CPU cores located on the same NUMA socket. Even so, during normal operations it is inevitable to access packets across NUMA socket boundaries: for instance, when an incoming packet is dispatched to a forwarding thread on another NUMA socket, or when the egress interface happens to be on a different NUMA socket.

2.2 Sharded Data Structures

A classical NDN forwarder has three main data structures (or *tables*): the FIB guides Interest forwarding toward the producer, the PIT gets Data back to the consumer, and the CS provides in-network caching. NDN-DPDK has multiple forwarding threads and they all need to access these three tables. Concurrent access, however, requires thread safety, which would increase the design complexity and reduce performance.

Instead, NDN-DPDK tries to avoid sharing the tables as much as possible: each forwarding thread has a private instance of the PIT and the CS, and a (partial) copy of the FIB. The first two do not require any cross-thread access, thus they are implemented using non-thread-safe data structures, while the FIB still needs to be updated from the control plane and therefore employs a low-overhead *Read-Copy-Update* (RCU) synchronization mechanism (section 4). This approach also enables allocating all three tables for each thread in NUMA-local memory in order to minimize the access latency, as explained in section 2.1.

2.3 Internal Packet Queues

Each forwarding thread receives the packets dispatched to it by the input stage via a set of three FIFO queues, one for each packet type: Interest, Data, Nack. The queues are provided by DPDK's *ring* library, which implements a ring buffer with a fixed capacity and lockless enqueue/dequeue operations. Packets are taken from any one of these queues in *bursts* instead of one at a time. Burst dequeuing amortizes the overhead of the ring buffer bookkeeping operations and reduces the number of CPU instruction cache misses, because multiple packets of the same type are processed together, typically following the same code path.

Initial testing suggested that the forwarding stage can easily become a bottleneck in configurations with few forwarding threads. To alleviate this bottleneck, NDN-DPDK adopts a CoDel-based [15] queue management algorithm between the input stage and the forwarding stage. If the minimum queuing delay stays above *target* (5 ms by default) for *interval* milliseconds (initially 100 ms), the forwarding thread inserts a *congestion mark* [25] in the next packet, prompting the consumer to slow down.

The forwarder also prioritizes Data over Interests, by dequeuing fewer packets from the Interest queue than from the Data queue at each iteration. This is because dropping a Data packet would not only waste the resources already spent on processing the corresponding Interest, but also cause the PIT entry to linger until its expiration, while losing an Interest is less harmful.

2.4 Life of a Packet

An incoming frame is received by an input thread, which parses it according to the NDN Packet format [19] and recognizes it as an Interest (section 3.1). The packet is then assigned to a forwarding thread based on its name (section 3.2) and is placed on a queue that goes to that thread. The forwarding thread dequeues the Interest. It first queries the PIT and the CS (section 5), but, assuming that this node did not recently receive any Interest or Data with the same name, no match is found in either table. Therefore, the thread creates a new PIT entry and records the ingress interface as a downstream node. Next, it performs a FIB lookup (section 4) to determine which forwarding strategy should be making forwarding decisions and the potential next hops. The strategy is invoked and decides to forward the Interest to a particular next hop (section 6.1). The forwarding thread finally passes the Interest to an output thread for transmission (section 3.4).

When a Data packet arrives, the input thread determines which forwarding thread previously handled the corresponding Interest based on a hop-by-hop header field in the Data packet (section 3.3), and passes the packet to that thread. The forwarding thread first locates the PIT entry that can be satisfied (section 5.2). It then checks which downstream nodes have expressed matching Interests and passes copies of the Data packet to output threads for transmission (section 3.4). It also notifies the forwarding strategy that the Interest has been satisfied, so that the strategy can make better decisions in the future (section 6). Finally, the Data packet is cached in the CS and the PIT entry is deleted.

3 FACE I/O AND PACKET DISPATCHING

NDN-DPDK is optimized for 10/100 Gbps Ethernet adapters attached to a PCIe bus. Using DPDK's *poll mode* drivers, NDN-DPDK can send and receive packets directly from user space without going through the operating system kernel. This significantly improves performance by eliminating the overhead of system calls and interrupt handling.

An NDN *face* is a generalization of the concept of network interface, on which NDN packets can be transmitted and received. NDN-DPDK supports only point-to-point faces, unlike other NDN forwarders, and deliberately does not support multicast communication on a single face. This directly follows from NDN-DPDK's primary use case in high-performance computing and data-intensive science [35], where the network is closely managed and peers are administratively configured, thus multicast-based discovery is not needed. Furthermore, to keep the face system simple, NDN-DPDK currently supports only Ethernet faces and cannot tunnel over IP or any other protocols[2]. The user can create one or more faces on a given Ethernet adapter; each face is distinguished by a different remote MAC address and optionally a VLAN ID.

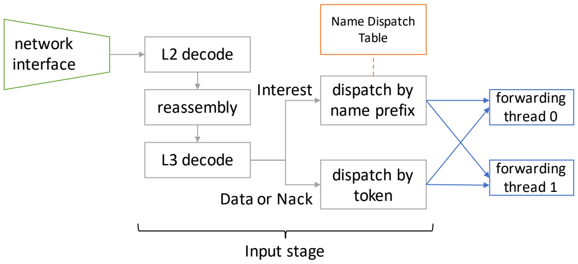


Figure 2: Input thread procedure.

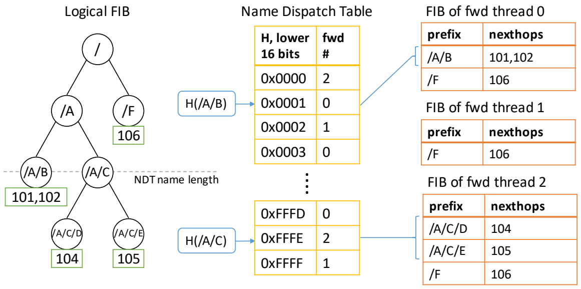
3.1 Input Stage

The forwarder's input pipeline, depicted in fig. 2, starts with receiving frames from an Ethernet adapter. As mentioned before, there can be multiple faces on the same adapter, but each face must have a distinct remote MAC address or VLAN ID. The adapter is configured to steer frames belonging to each face into a different receive queue. An *input thread* is connected to one or more of these receive queues and is responsible for decoding each incoming frame with a three-step procedure: (1) strip the Ethernet header and any VLAN tags; (2) decode as an NDNLP packet [21], performing reassembly if needed; (3) continue decoding the reassembled NDN network-layer packet and classify it as Interest, Data, or Nack.

The decoding routines store information about a parsed packet along with the packet buffer itself, in a private area reserved in the buffer header. This avoids the need for a separate memory allocation, which improves throughput but consumes more memory on a per-packet basis. Moreover, NDN-DPDK's decoding routines are optimized to serve the forwarder and do not have to accommodate the needs of consumer or producer applications. This leads to several design differences compared to the parsers that can be found in any general-purpose NDN library. For example:

- For Data, decoding stops just before the `Content` element, because the forwarding algorithms do not need to know the packet's payload or its signature.
- For similar reasons, the decoder ignores an Interest's `ApplicationParameters` element and everything that comes after it.
- For the `MetaInfo` element in a Data packet, the decoder reads only the `FreshnessPeriod` field, because other fields do not affect forwarding.

[2] Better tunneling support is planned for a future version.



H, lower 16 bits	fwd #
0x0000	2
0x0001	0
0x0002	1
0x0003	0
⋮	
0xFFFD	0
0xFFFE	2
0xFFFF	1

prefix	nexthops
/A/B	101,102
/F	106

prefix	nexthops
/F	106

prefix	nexthops
/A/C/D	104
/A/C/E	105
/F	106

Figure 3: Name Dispatch Table and FIB partitions.

3.2 Dispatching Interests by Name

After packet decoding, the input thread dispatches each network-layer packet to a forwarding thread. For an Interest, a prefix of the name determines which forwarding thread will process the packet. This dispatching algorithm is name-based so that two Interests with the same name end up in the same forwarding thread, which ensures the effectiveness of Interest aggregation. By using a name prefix instead of the entire name, Interests with a common prefix go to the same forwarding thread, which enables the forwarding strategy to collect measurements on a per-prefix basis.

As we will see in section 6, forwarding strategies operate at the FIB entry granularity. In theory, Interest dispatching could follow the FIB entries, which would fulfill the two goals above. However, the FIB is a fairly complex data structure and performing a full lookup in the input stage would be too slow. Therefore, the dispatching algorithm employs a much simpler method to determine the granularity: it takes the first k components of the Interest name as prefix. If the Interest has fewer than k name components, the entire name is used. The value of k should be chosen such that the resulting prefixes are shorter than most FIB entries, to keep forwarding strategy measurements effective, but also long enough that there is a sufficient number of distinct prefixes for load balancing among forwarding threads. NDN-DPDK sets $k = 2$ in its default configuration, but this parameter can be tuned according to the characteristics of the incoming traffic.

After determining the name prefix, the dispatching algorithm queries the *Name Dispatch Table* (NDT), a table unique to the NDN-DPDK forwarder (fig. 3). The input thread dispatches the Interest to the queue leading to the forwarding thread identified by the NDT entry. In order to be as simple as possible and maintain a predictable lookup speed, the NDT is not a name-indexed data structure but a linear vector of 2^b entries ($b = 16$ by default), where each entry contains a forwarding thread identifier. The lookup algorithm computes the SipHash [2] over the name prefix, takes the lower b bits of the hash value as an index into the vector, and returns the forwarding thread identifier in that entry. This allows an NDT lookup to complete in $O(1)$ time complexity.

3.3 Dispatching Data by Token

Data must be dispatched to the same forwarding thread that processed the Interest. As described in section 2.2, each forwarding thread has a private instance of the PIT. Therefore, information about a pending Interest is available only in the forwarding thread that forwarded the Interest and only that thread is able to correctly process a Data in reply to the forwarded Interest.

Although we could perform another prefix-based dispatch on the Data name, the name dispatching algorithm breaks down in one specific case. The NDN protocol allows name discovery: an Interest may be satisfied if its name is a prefix of the Data name and the Interest contains the `CanBePrefix` flag. For example, Data `/A/B/1` can satisfy an Interest with name `/A` and `CanBePrefix=1`. Applying the name dispatching algorithm in section 3.2 and observing that the Interest name in this case has fewer than two components, we can see that the Data would be dispatched under the `/A/B` prefix, while the Interest would be dispatched under the `/A` prefix. If the NDT entries corresponding to these prefixes map to two different thread identifiers, the Data would go to a forwarding thread that has no knowledge about the Interest.

To solve this problem we introduce the *PIT token*, an 8-byte hop-by-hop field carried in the NDNLP header of each packet [21]. Downstream nodes attach a PIT token when transmitting an Interest. Upstream nodes are expected to attach the same token when replying to the Interest with a Data or Nack packet. NDN-DPDK uses a few bits of this token to encode the forwarding thread identifier. When a forwarding thread transmits an Interest, it puts its own identifier in the PIT token. When Data comes back, the input thread can simply read that part of the PIT token and dispatch the Data to the correct forwarding thread.

For a Nack packet, either dispatching method would work correctly. We chose the token dispatching method as it is more efficient.

3.4 Output Stage

The output stage controls the transmission of outgoing packets. For each outgoing network-layer packet, it performs NDNLP fragmentation (if needed), prepends an Ethernet header, and enqueues the resulting link-layer frames for transmission on the Ethernet adapter. In the current version of NDN-DPDK the workload of the output threads is light. However, it will likely increase in the future as we plan to implement more advanced congestion control and queue management schemes in the output stage.

4 FIB STRUCTURE AND LOOKUP

The FIB is a read-mostly table. The forwarding threads perform a *Longest Prefix Match* (LPM) on the FIB to determine where to send each incoming Interest. On the other hand, FIB updates are seldom needed and only occur in response to a management command from the control plane.

NDN-DPDK's FIB design is inspired by So et al. [30]. In their design, the FIB entries are stored in a hash table keyed by the name prefixes. They also propose a 2-stage LPM lookup algorithm:

(1) The FIB has a fixed parameter M, such that the majority of FIB entry names have fewer than M components.
(2) When inserting a FIB entry whose name has more than M components, a virtual entry is inserted at depth M that indicates the maximum FIB depth under its prefix.
(3) Given an Interest, an LPM on the FIB starts with the M-component prefix of the Interest name.

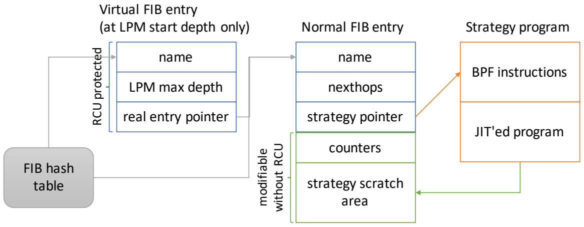


Figure 4: Structure of normal and virtual FIB entries.

(4) If this lookup finds a virtual entry, the LPM is restarted at the maximum FIB depth indicated in the virtual entry.
(5) Otherwise, the LPM continues toward shorter prefixes until a normal FIB entry is found.

In NDN-DPDK, we adopted the same FIB structure (fig. 4) and the same 2-stage lookup algorithm. However, our multi-threaded architecture requires thread safety, which we achieved with two techniques.

First, we use the Userspace Read-Copy-Update (URCU) library [3, 13] to allow both FIB queries from the forwarding thread and FIB updates from the management thread to take place simultaneously. A benefit of RCU is that its read-side overhead is minimal, which matches well with the read-mostly nature of the FIB. We use the quiescent-state-based flavor of RCU because it has the smallest overhead. This flavor requires every thread to periodically indicate a quiescent state; for a forwarding thread, this occurs before processing each burst of packets. The hash table implementation comes from URCU's lock-free resizable RCU hash table, with the resize functionality disabled to provide more predictable performance. FIB entries are allocated from a DPDK memory pool instead of the default `malloc()` memory allocator.

Second, each forwarding thread is given its private FIB instance. Each FIB instance contains only the name prefixes served by the forwarding thread. Compared to having a single FIB shared among all forwarding threads, this approach ensures the FIB entries are allocated on the same NUMA socket as the forwarding thread, avoiding memory access across NUMA boundary. Moreover, it allows the forwarding strategy (section 6) to store collected measurements on the FIB entry itself, without needing a separate measurements table.

5 COMBINED PIT AND CONTENT STORE DESIGN

The PIT and the Content Store are both read and modified frequently in the data plane, specifically:

(1) The CS is queried for every incoming Interest to check if it can be satisfied by cached Data.
(2) If not, the Interest is forwarded and a PIT entry must be inserted, unless one already exists.
(3) Upon receiving a Data packet, the PIT is queried to find which pending Interest(s) can be satisfied.
(4) When a PIT entry is satisfied, it must be erased and a CS entry inserted to cache the Data.
(5) If the CS is full, it may need to evict some entries to make room for the new Data.

Observing that item 4 often deletes a PIT entry and inserts a CS entry at the same name prefix, So et al. [30] propose combining the PIT and the CS into a single hash table, so that a satisfied PIT entry can be replaced with a CS entry without incurring the cost of a second table lookup. NDN-DPDK adopts this design and merges PIT and CS into the *PIT-CS Composite Table* (PCCT).

However, [30] is intended for the CCNx protocol [14], which differs from NDN in the Interest-to-Data matching rules. Both NDN and CCNx allow a Data[3] packet to satisfy an Interest if they have the same name. NDN additionally allows a Data to satisfy an Interest if the Interest name is a prefix of the Data name and the Interest carries the `CanBePrefix` flag. Moreover, NDN Interests can carry a forwarding hint that, when present, should be used in place of the Interest name to determine the forwarding path. These major protocol differences make our PCCT design inevitably different from the one described in [30].

Given the frequent updates in both PIT and CS, a thread-safe PCCT shared across all forwarding threads could easily become a bottleneck. We decided early on that each forwarding thread should have its own private instance of the PCCT. Contrary to the FIB case, the control plane does not need to interact with either the PIT or the CS. Therefore, we can implement the PCCT using non-thread-safe data structures, which are typically faster than their thread-safe counterparts.

5.1 Logical Structure

The overall structure of the PCCT is a combination of three data structures:

- A DPDK *mempool* to allocate PCCT entries from.
- A *name hash table* for name-based lookups. This reuses the SipHash values already computed during packet dispatching (section 3.2).
- A *token hash table* to find what PIT entries can be satisfied by incoming Data, using the PIT token carried on the Data packet (section 5.2).

Logically, each PCCT entry contains a name, a chosen forwarding hint, two PIT entries, and one CS entry (fig. 5). The name is required, but all other fields are optional.

An entry is identified by the combination of its name and the chosen forwarding hint. When a forwarding thread processes an Interest that carries forwarding hints, it performs FIB lookups using those hints, and chooses the first hint that matches a FIB entry. The PIT entry created from that Interest and the CS entry for its reply Data are then placed on a PCCT entry with the chosen forwarding hint. Having the latter as part of the PCCT entry identifier logically isolates the PIT and the CS for each forwarding hint into different partitions. This mitigates a well-known cache poisoning attack caused by forwarding hints and makes NDN-DPDK the first NDN forwarder to *securely* support forwarding hints.

Each PCCT entry can contain up to two PIT entries with the same name and chosen forwarding hint. Per the NDN protocol [19], an Interest may carry the `CanBePrefix` and/or `MustBeFresh` flags

[3] "Data" is NDN terminology; the CCNx equivalent is "Content Object".

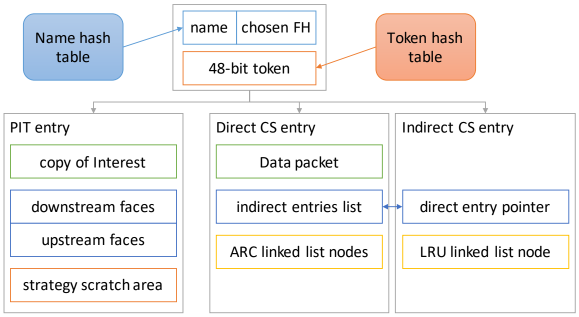


Figure 5: PIT-CS Composite Table.

that affect Interest-to-Data matching in forwarding and caches. However, the protocol is vague on how the PIT should aggregate Interests with the same name but different flags. We argue that two Interests that differ only in the `CanBePrefix` flag can be aggregated because the presence of this flag widens the set of Data that can be matched. On the other hand, two Interests that differ in the `MustBeFresh` flag cannot be aggregated, because forwarding them with the flag set would reject non-fresh Data that would otherwise satisfy the Interest that did not have the flag, while forwarding them without the flag could incorrectly satisfy the `MustBeFresh` Interest with non-fresh Data. Hence, we decided to store up to two PIT entries in each PCCT entry, one with `MustBeFresh` and one without.

5.2 PIT Lookup by Data

As mentioned before, the NDN protocol allows prefix match between Interest and Data. When a Data packet arrives, the forwarder would have to perform an LPM lookup on the PIT to determine which pending Interests can be satisfied. This could become a performance bottleneck or even an attack surface because the Data name can have many components. It is infeasible to apply the 2-stage lookup algorithm (section 4) to the PIT because the overhead of maintaining M would be too high. Instead, we propose a different approach based on the *PIT token*, a short hop-by-hop header field already introduced in section 3.3 for Data packet dispatching. Here, we extend its usage to accelerate PIT lookups.

Whenever a forwarding thread inserts a PIT entry, it allocates a *PCCT entry token* to the enclosing PCCT entry and adds it to the token hash table. Then, every outgoing Interest created from this PIT entry will carry the PCCT entry token inside its PIT token field. When a Data packet comes back, the forwarding thread can quickly locate the PCCT entry in $O(1)$ time via the token hash table. It is still necessary to verify that the Data indeed satisfies the Interest through name comparison, to prevent attacks from forged tokens.

5.3 Prefix Matching in the Content Store

Being a hash table, the PCCT only supports exact match queries using a key that consists of an Interest/Data name and a chosen forwarding hint. Thus, if an Interest name is a prefix of the Data name and the Interest carries the `CanBePrefix` flag, the exact match algorithm cannot retrieve the Data from the hash table. This severely limits the effectiveness of in-network caching, because any Interest carrying a non-exact name would have to be answered by the producer instead of being satisfied from the CS.

To address this issue, NDN-DPDK introduces the concept of *indirect CS entry* to provide partial support for prefix matching. An indirect CS entry is a special CS entry named after an Interest and containing a pointer to a *direct CS entry*. By contrast, a direct CS entry is a regular CS entry named after the Data, and contains the bits of the Data packet itself.

When the forwarder receives a Data in reply to an Interest with non-exact name, it inserts two entries into the CS: a direct entry with the Data name and the Data packet, and an indirect entry with the Interest name and a pointer to the direct entry. This allows CS matching with a prefix name, under the assumption that the consumer application consistently uses the same prefix (or a small subset of prefixes) to perform name discovery. If a future Interest with the same name as the previous Interest arrives, an exact match lookup in the CS using that Interest name will find the indirect entry, from where we can follow the pointer and retrieve the direct entry and the cached Data packet. Conversely, if an Interest with a different name arrives, even if it matches the Data, the forwarder will not be able to find the cached Data because an indirect CS entry for that Interest name does not exist.

6 FORWARDING STRATEGIES

The forwarding strategy is a component that controls various aspects of the Interest forwarding behavior. The strategy decides where to forward an incoming Interest when it cannot be satisfied by the local CS. It also decides whether to perform any corrective actions when a Nack is received. These decisions are based on inputs such as the next hops in the matching FIB entry, the downstream and upstream records in the PIT entry, as well as any collected measurements on recent data retrievals by Interests sharing a common prefix.

Experience with early NDN deployments has shown that different applications need different Interest forwarding behaviors. This provides a strong motivation to support multiple forwarding strategies with different decision making algorithms, and to dynamically choose a forwarding strategy based on application needs and network environments. As outlined by Jacobson et al. [9], the basic idea is for each FIB entry to contain a program, written for an abstract machine specialized to forwarding choices, that determines how to forward Interests. NDN-DPDK realizes this vision using *extended BPF* (eBPF) [5, 18], an evolution of the original Berkeley Packet Filter (BPF) [12].

6.1 Strategy Program

Each strategy is an eBPF program, i.e., a list of low-level instructions, such as load/store, arithmetic, and comparison operators, that can be executed by the eBPF virtual machine. In addition, the program can call a few higher-level routines that are provided to the strategies by the core forwarding engine. These include setting a timer, sending the current Interest on the specified face, and responding to an Interest with a Nack.

The strategy program exports a main function that is invoked, or *triggered*, when:

- An Interest packet arrives and cannot be satisfied with cached Data. The strategy will have to decide how to forward it.
- A Data packet arrives and satisfies an Interest. This trigger allows the strategy to collect path measurements, such as round-trip time and satisfaction ratio.
- A Nack packet arrives and does not fall under one of the simple cases that are automatically handled by the forwarding plane itself. The strategy can then decide if any corrective actions must be taken.
- A timer previously scheduled by the strategy expires.

Both FIB and PIT entries contain a writable *scratch area* for the strategy to store its state and record measurements. The FIB entry scratch area (fig. 4) is suitable for information related to a whole namespace, while the PIT entry scratch area (fig. 5) is suitable for information related to a specific pending Interest. Other than these areas, the strategy is able to inspect the current packet and has read-only access to a subset of fields in the FIB and PIT entries.

6.2 Strategy Selection and FIB Updates

NDN-DPDK associates a forwarding strategy to every FIB entry. When an Interest arrives and cannot be satisfied by the CS, the forwarding plane performs a FIB lookup, and the matched FIB entry determines not only the potential next hops but also which strategy should handle the Interest. This design is in line with [9], except that the strategy eBPF program is not stored in the FIB entry itself, but referenced by the FIB entry, so that the same strategy may be used by multiple FIB entries without storing duplicate copies. Data and Nacks are always handled by the same strategy that processed the Interest. In case the strategy or the FIB entry was changed while the Interest was pending, the forwarder handles the returning Data/Nack packet using a built-in fallback procedure and no strategy is triggered.

After a FIB update, the strategy has to restart with an empty scratch area, because it would be infeasible to migrate the data in the scratch areas during FIB updates. Indeed, although FIB entries are RCU-protected (section 4), the FIB entry scratch area is not. While this allows the strategy to modify the scratch area without going through the relatively expensive write-side RCU procedure, it also means that the control plane thread cannot safely copy the scratch area contents from the old FIB entry to the new one.

7 PERFORMANCE EVALUATION

We conducted extensive testing of NDN-DPDK in order to determine its performance characteristics under a variety of workloads. In particular, we measured the *aggregate forwarding rate*, in terms of bps (bits per second) and pps (Data packets per second), and the *per-packet forwarding latency*, in microseconds. Note that our pps metric accounts only for the Data packets because they are carrying the application content. The total number of packets (Interests and Data) actually forwarded by NDN-DPDK is at least twice[4] the reported amounts.

In all the experiments described below, the forwarder is running on a Supermicro 6039P-TXRT server equipped with dual Intel Xeon Gold 6240 CPUs (18 cores at 2.60 GHz, with Hyper-Threading disabled), 256 GB of 2933 MHz memory in four channels (64 × 1 GB hugepages have been allocated to NDN-DPDK on each NUMA socket), and Mellanox ConnectX-5 100 Gbps Ethernet adapters. The operating system is Ubuntu Linux 18.04, with DPDK v19.11 and NDN-DPDK commit 34f561f4ef0e5790d4999107dcbb4c2eab82af66. The forwarder node is connected to two traffic generators, one on each Ethernet port, via direct attach copper cables. The traffic generators emulate a producer application and a number of consumers requesting content from the producer. Each consumer instance expresses Interests under a given name prefix and employs a congestion control algorithm similar to TCP CUBIC [24]. The Interest names consist of five distinct parts:

1. The producer prefix.
2. The consumer thread ID.
3. The consumer node name followed by a random number that changes with each execution.
4. The placeholder component `/127=Y` repeated as many times as necessary (possibly zero) to make the total Interest name length equal that required by the experiment scenario.
5. The segment number.

An example Interest name is `/C/0/B_77378826/127=Y/127=Y/35=%07%C3`. The producer responds to each Interest with a Data packet, either of the same name or with an additional suffix name component `/127=Z`. The Data packet signature is neither generated by the producer nor verified by the consumer, although a signature field of proper length but with a fictitious value is present in every Data packet.

All experiments share a common "base configuration" consisting in: 8 forwarding threads, 4 components in Interest and Data names, 1000 bytes of application payload in every Data packet, 2^{16} NDT entries (see section 3.2), FIB start depth (the M parameter in section 4) set to 8, and a maximum CS capacity equal to 2^{15} entries per forwarding thread (see section 5). In each experiment we vary some of these parameters in order to assess their impact on the overall performance of the forwarder.

The forwarding rate reported for each benchmark is the arithmetic mean of 10, 50, or 150 consecutive runs (depending on the experiment), executed after a warm-up run whose results are discarded. Each run lasts 60 seconds. The forwarding latency is measured for each packet, from the moment it enters the forwarder's input stage to when it is dequeued by the output thread and handed over to the network adapter for transmission.

7.1 Forwarding Threads and Name Length

This benchmark demonstrates how NDN-DPDK's performance scales with the number of CPU cores assigned to the packet forwarding tasks. As described in section 2, each CPU core used by NDN-DPDK is entirely dedicated to running one and only one thread, hence we will use the terms "core" and "thread" interchangeably. Given that the number of input and output threads is constrained by how many network cards are installed on the system, we can only vary the number of forwarding threads, in order to distribute the incoming traffic among a larger or smaller amount of CPU cores.

[4] We say "at least twice" and not "exactly twice" because in the rare case of packet loss, the consumer must retransmit the Interest. These retransmissions are not included in our statistics since they do not affect the final application-layer goodput.

Shi, Junxiao; Pesavento, Davide; Benmohamed, Lotfi. "NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware." Paper presented at 7th ACM Conference on Information-Centric Networking (ICN 2020), Montreal, CA. September 29, 2020 - October 01, 2020.

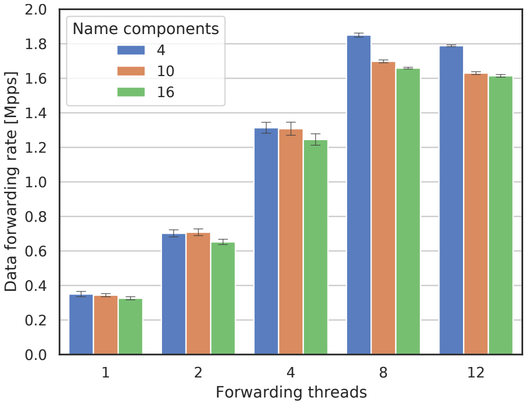


Figure 6: Mean and standard deviation of the throughput with different numbers of forwarding threads and name components.

Figure 6 shows that the throughput grows almost linearly up to 4 forwarding threads, then slows down but still improves up to 8 threads, where we reach a peak rate of about 1.84 Mpps (million Data packets per second). Further increasing the number of threads beyond that does not help performance, in fact the throughput slightly declines with 12 threads, as more cores on the same CPU are competing for hardware resources. A deeper analysis, not reported here for lack of space, revealed that with 8 or more forwarding threads, the input stage of the forwarder's pipeline becomes the bottleneck. We plan to eliminate this bottleneck in a future version of NDN-DPDK.

In fig. 6 we also illustrate the effect of the Interest name length, expressed in number of name components, on the overall forwarding rate. Longer names make the forwarder marginally slower, up to 10 % in the worst case, and the slowdown is more pronounced with 8 and 12 forwarding threads, i.e., when the input stage becomes the bottleneck. This can be explained by the fact that it is the input thread that parses the Interest name (section 3.1) and the time complexity of the decoding algorithm is linear in the number of name components.

7.2 Data Payload Length

The application-layer payload is carried inside NDN Data packets in a field called `Content`. As detailed in section 3.1, NDN-DPDK never decodes or otherwise accesses this field, because it is completely opaque to NDN routers and does not affect any forwarding decision. Therefore, we expect that varying the Data payload length will have very limited impact on the forwarder's performance. The benchmark results in fig. 7 indeed confirm our intuition: the difference between best-case and worst-case pps forwarding rate never exceeds 10 %.

This experiment also allows us to determine the maximum aggregate throughput in bits per second (bps) that can be delivered by NDN-DPDK between two network adapters: 108 Gbps, with 8000 bytes of content per packet. This number represents the application-layer *goodput*, thus excluding all Ethernet headers, Interests/Data names, Data signatures, and so on.

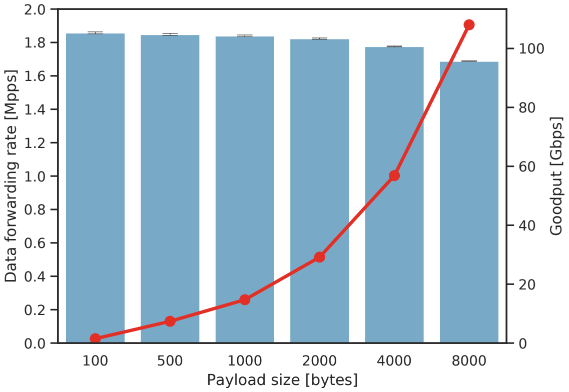


Figure 7: Mean and standard deviation of the throughput in Mpps (bars) and Gbps (line) with varying Data payload sizes.

7.3 Scalability of FIB Lookup

We tested the scalability of NDN-DPDK's FIB lookup algorithm (section 4) with up to 1 million entries (name prefixes). We can notice from the table below that the number of FIB entries has no impact on the forwarding rate, while the Interest latency is only minimally affected.

	Fwd. rate (kpps)		Interest latency (µs)	
FIB entries	Mean	σ	Median	95th percentile
10^4	1840	5.59	90	227
10^5	1835	4.92	92	234
10^6	1839	4.42	97	249

7.4 Content Store Capacity and Hit Ratio

One major feature of NDN-DPDK is the built-in support for a limited form of prefix matching in the Content Store, i.e., the ability to return, in some cases, a cached Data packet that has a longer name than the incoming Interest. However, this type of lookup is more expensive than a simple exact match algorithm due to the additional indirection and the greater number of PCCT entries that need to be consulted, as explained in section 5.3. With this benchmark we tried to quantify the performance difference between exact and prefix match. In order to trigger a meaningful number of CS hits, we added a second consumer node to the experiment topology, connected to the forwarder through a third Ethernet port. The second consumer starts fetching content 100 ms after the first one.

Figure 8 shows the aggregate throughput results of 150 runs, together with the linear regression line obtained via Theil-Sen estimation [26, 32]. Overall, prefix matching is 15 % to 17 % slower than exact matching when the CS capacity is set to 2^{17} entries per thread, and between 10 % and 18 % slower with a capacity of 2^{20} entries. We can also see from the figure that the forwarding rate is positively correlated with the hit ratio. This is because satisfying

Shi, Junxiao; Pesavento, Davide; Benmohamed, Lotfi. "NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware." Paper presented at 7th ACM Conference on Information-Centric Networking (ICN 2020), Montreal, CA. September 29, 2020 - October 01, 2020.

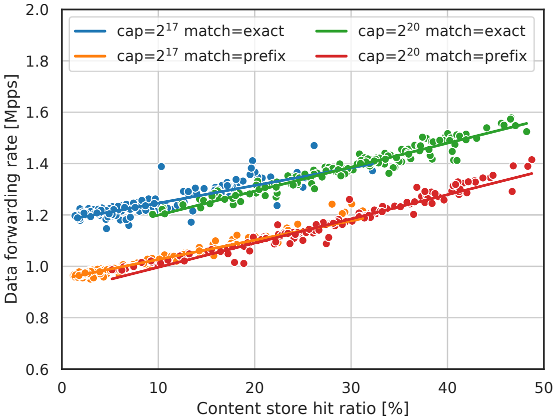


Figure 8: Forwarding rate vs. hit ratio with exact and prefix match and different Content Store capacities.

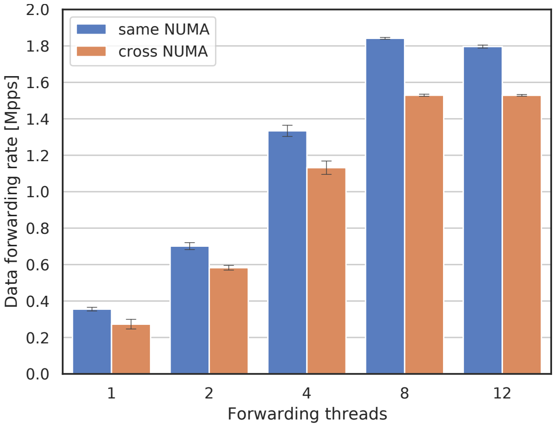


Figure 9: Mean and standard deviation of the throughput under same-NUMA and cross-NUMA memory accesses.

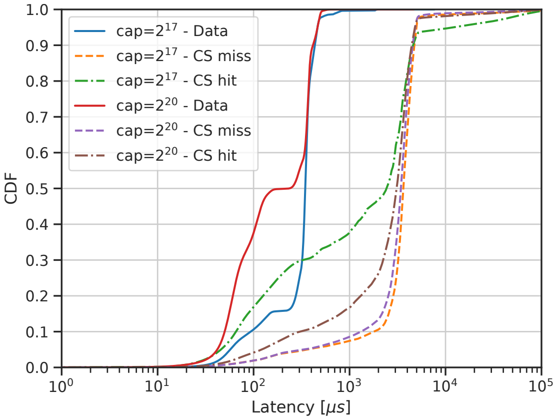


Figure 10: Cumulative distribution function of the forwarding latency with exact name matching.

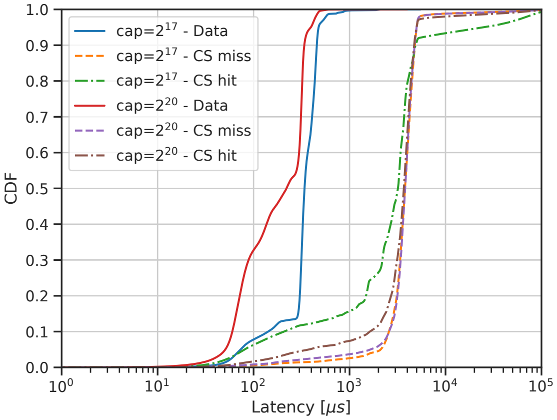


Figure 11: Cumulative distribution function of the forwarding latency with prefix name matching.

more Interests from the cache means that fewer packets will have to be further processed through the pipeline stages.

We also looked at the processing latency of each of the three main code paths that an incoming packet can take: *CS miss* (an Interest that is forwarded upstream), *CS hit* (an Interest that is answered with a Data packet from the local cache), and *Data* (a Data packet that is inserted into the CS and then forwarded downstream). The latency distributions are plotted in figs. 10 and 11 for the two scenarios of exact and prefix matching.

7.5 Impact of Nonlocal Memory Access

As mentioned in section 2.1, accessing memory on another NUMA socket incurs a higher latency compared to accessing local memory. In this benchmark, we measured the impact of cross-NUMA memory accesses on NDN-DPDK's performance. We ran the same test twice: first between two network cards installed on the same NUMA node as the forwarding threads, then we moved one of the cards to the other NUMA node, thereby forcing all traffic to require non-local memory accesses in either the Interest or the Data direction. The experiment was repeated 50 times for each scenario. The results in fig. 9 confirm that repeatedly crossing the NUMA boundary can significantly degrade the packet processing throughput, up to 20 % slower with 8 forwarding threads.

8 RELATED WORK

The research community has studied numerous performance aspects of the NDN and CCN forwarding models in great detail [11, 31, 33, 34, 36–38]. However, most of these works focus on just one or few facets of the problem, such as scalable FIB lookup, efficient name encoding, distributed PIT architectures, and hierarchical Content Stores, but only a handful of papers propose a complete design for a high-speed name-based forwarding engine.

Shi, Junxiao; Pesavento, Davide; Benmohamed, Lotfi. "NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware." Paper presented at 7th ACM Conference on Information-Centric Networking (ICN 2020), Montreal, CA. September 29, 2020 - October 01, 2020.

Caesar [16] is the first full implementation of a content-centric router with a design based on hash tables. Its data plane runs on a specialized hardware platform consisting of four 10 GbE line cards and takes advantage of several hardware-specific accelerations. The FIB is distributed across the line cards and, in addition to a hash table, it uses a prefix Bloom filter, which requires hardware assistance to be beneficial. *Caesar* does not support prefix matching between Interest and Data names and relies on a custom packet format with a few fixed-length header fields to expedite parsing, therefore it cannot easily be extended to handle NDN semantics.

Another notable router design centering on hash tables is proposed by So et al. [30]. NDN-DPDK took several key ideas from this paper, enhancing them to provide additional features and support the more powerful NDN name matching semantics, as explained in the previous sections. Among them, the 2-stage LPM algorithm, the PIT partitioning scheme, and the unification of the PIT and CS data structures. A similar FIB lookup scheme is also described by Fukushima et al. [6]; however, this approach is effective only on FIB entries that are leaf nodes in the name hierarchy tree.

Kirchner et al. [10] present two open-source implementations of their software CCN router *Augustus*: a standalone monolithic forwarding engine based on DPDK and running on general-purpose hardware, and a modular prototype built with the Click framework. Their solution is able to reach a throughput of 10 Mpps, but it suffers from several shortcomings that preclude its practical deployment in NDN networks. For instance: (1) It implements a custom ICN packet format that contains a fixed-length header that complicates future evolutions of the protocol, in violation of NDN's "universality" design principle [20]. (2) Similarly to *Caesar*, it can perform only an exact match between Interest and Data packets. (3) To take advantage of hardware dispatching, *Augustus* encapsulates ICN packets in IPv4 packets and configures the adapter's Receive Side Scaling (RSS) in such a way that the hash result depends only on the source IPv4 address. To ensure that a returning Data packet is dispatched to the thread that has the corresponding PIT entry, *Augustus* requires the IPv4 source address field to contain the CRC32 hash of the name. This technique, while functionally similar to NDN-DPDK's PIT token, requires neighboring routers to agree on a hash function before they can communicate. (4) The FIB is not designed to be thread-safe, thus rendering FIB updates impossible while the forwarder is handling data plane traffic.

Other approaches that use hop-by-hop state carried between Interest and Data to accelerate or eliminate PIT lookups have been proposed in CCN-GRAM [7] and ADN [8].

9 CONCLUSION AND FUTURE WORK

This paper describes NDN-DPDK, our implementation of a high-performance NDN forwarder on commodity server hardware. NDN-DPDK is built upon a number of novel ideas, such as the introduction of the PIT token for efficient Interest-Data prefix matching, the introduction of indirect CS entries for efficient CS prefix matching, resulting in more effective in-network caching, and the secure support for NDN's forwarding hints. Initial benchmarks demonstrate that NDN-DPDK is capable of sustaining 1.8 Mpps, or a corresponding 108 Gbps when 8 kB Data packets are used.

Lessons learned from the benchmarking effort helped us identify future work needed for performance improvements in NDN-DPDK. As we observed, the input thread becomes the bottleneck when there are 8 or more forwarding threads. We are exploring potential design changes to allow attaching multiple input threads to the same Ethernet adapter. At the same time, we hope that hardware acceleration can speed up packet dispatching by sending the vast majority of incoming packets directly to a forwarding thread, bypassing the input thread bottleneck and ensuring that the packet buffers are allocated from a memory pool on the same NUMA socket as the forwarding thread. This goal can be achieved progressively:

1. If the network card's RSS supports matching at arbitrary offsets in the Ethernet frame, Data and Nack can be dispatched directly, by configuring RSS to read the PIT token field. This will reduce the work of the input thread to just processing Interests and fragmented packets.
2. Dispatching Interests will likely require eBPF/P4 hardware or FPGA-based solutions. One idea is to download a copy of the NDT into the hardware accelerator, which will then (partially) decode the incoming NDN packets and perform NDT lookups with the Interest names, using the algorithm in section 3.2. At this point, only fragmented packets need to be processed by software input threads.

Our roadmap for future releases also includes: (1) Expanding the Content Store capacity by caching some packets in persistent memory (e.g., Intel Optane) or NVMe disk storage. These devices are more cost effective and energy efficient than traditional RAM, and would potentially allow a forwarder to have caching capacity in the order of a few terabytes. However, their slower access speed requires novel multi-tier caching algorithms. (2) Automatic load balancing among forwarding threads by dynamically adjusting the NDT entries to spread the forwarder's workload. (3) VXLAN tunnelling with hardware offloads. (4) Continuous in-depth performance profiling, to guide further optimizations in memory access (e.g., CPU cache prefetching) and data structure design (e.g., hash table collision resolution algorithms). (5) Implementation and benchmarking of in-forwarder cryptographic processing, such as implicit digest for Data packets. (6) New forwarding strategies with enhanced capabilities.

DISCLAIMER

Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST, nor does it imply that the products mentioned are necessarily the best available for the purpose.

REFERENCES

[1] Mohammad Alhowaidi, Byrav Ramamurthy, Brian Bockelman, and David Swanson. 2017. The case for using content-centric networking for distributing high-energy physics software. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2571–2572.

[2] Jean-Philippe Aumasson and Daniel J Bernstein. 2012. SipHash: a fast short-input PRF. In *International Conference on Cryptology in India*. Springer, 489–508.

[3] Mathieu Desnoyers and Paul E. McKenney. [n.d.]. *Userspace RCU Implementation.* Retrieved August 31, 2020 from https://liburcu.org/

[4] Chengyu Fan, Susmit Shannigrahi, Steve DiBenedetto, Catherine Olschanowsky, Christos Papadopoulos, and Harvey Newman. 2015. Managing scientific data with named data networking. In *Proceedings of the Fifth International Workshop on Network-Aware Data Management.* 1–7.

Shi, Junxiao; Pesavento, Davide; Benmohamed, Lotfi. "NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware." Paper presented at 7th ACM Conference on Information-Centric Networking (ICN 2020), Montreal, CA. September 29, 2020 - October 01, 2020.

[5] Matt Fleming. 2017. A thorough introduction to eBPF. *LWN.net* (2017). https://lwn.net/Articles/740157/

[6] Masaki Fukushima, Atsushi Tagami, and Toru Hasegawa. 2013. Efficiently looking up non-aggregatable name prefixes by reducing prefix seeking. In *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 340–344.

[7] JJ Garcia-Luna-Aceves and Maziar Mirzazad Barijough. 2016. Content-centric networking using anonymous datagrams. In *2016 IFIP Networking Conference (IFIP Networking) and Workshops*. IEEE, 171–179.

[8] José Joaquin Garcia-Luna-Aceves. 2017. ADN: An information-centric networking architecture for the Internet of Things. In *Proceedings of the second international conference on internet-of-things design and implementation*. 27–36.

[9] Van Jacobson, Diana K Smetters, James D Thornton, Michael F Plass, Nicholas H Briggs, and Rebecca L Braynard. 2009. Networking named content. In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*. 1–12.

[10] Davide Kirchner, Raihana Ferdous, Renato Lo Cigno, Leonardo Maccari, Massimo Gallo, Diego Perino, and Lorenzo Saino. 2016. Augustus: a CCN router for programmable networks. In *Proceedings of the 3rd ACM Conference on Information-Centric Networking*. 31–39.

[11] Rodrigo B Mansilha, Lorenzo Saino, Marinho P Barcellos, Massimo Gallo, Emilio Leonardi, Diego Perino, and Dario Rossi. 2015. Hierarchical content stores in high-speed ICN routers: Emulation and prototype implementation. In *Proceedings of the 2nd ACM Conference on Information-Centric Networking*. 59–68.

[12] Steven McCanne and Van Jacobson. 1993. The BSD Packet Filter: A New Architecture for User-level Packet Capture. In *USENIX winter*, Vol. 46.

[13] Paul E. McKenney, Mathieu Desnoyers, and Lai Jiangshan. 2013. User-space RCU. *LWN.net* (2013). https://lwn.net/Articles/573424/

[14] M. Mosko, I. Solis, and C. Wood. 2019. Content-Centric Networking (CCNx) Semantics. RFC 8569 (Experimental). https://doi.org/10.17487/RFC8569

[15] K. Nichols, V. Jacobson, A. McGregor (Ed.), and J. Iyengar (Ed.). 2018. Controlled Delay Active Queue Management. RFC 8289 (Experimental). https://doi.org/10.17487/RFC8289

[16] Diego Perino, Matteo Varvello, Leonardo Linguaglossa, Rafael Laufer, and Roger Boislaigue. 2014. Caesar: A content router for high-speed forwarding on content names. In *Proceedings of the tenth ACM/IEEE symposium on Architectures for networking and communications systems*. 137–148.

[17] DPDK Project. [n.d.]. *Data Plane Development Kit*. Retrieved August 31, 2020 from https://www.dpdk.org/

[18] IO Visor Project. [n.d.]. *eBPF: extended Berkeley Packet Filter*. Retrieved August 31, 2020 from https://www.iovisor.org/technology/ebpf

[19] Named Data Networking Project. [n.d.]. *NDN Packet Format Specification, version 0.3*. Retrieved August 31, 2020 from https://named-data.net/doc/NDN-packet-spec/0.3/

[20] Named Data Networking Project. [n.d.]. *NDN Protocol Design Principles*. Retrieved August 31, 2020 from https://named-data.net/project/ndn-design-principles/

[21] Named Data Networking Project. [n.d.]. *NDNLPv2: NDN Link Protocol, version 2*. Retrieved August 31, 2020 from https://redmine.named-data.net/projects/nfd/wiki/NDNLPv2

[22] Named Data Networking Project. 2018. *NFD Developer's Guide*. Technical Report. NDN-0021, Revision 10. https://named-data.net/publications/techreports/ndn-0021-10-nfd-developer-guide/

[23] Duncan Rand, Simon Fayer, and David J Colling. 2015. Possibilities for named data networking in HEP. In *Journal of Physics: Conference Series*, Vol. 664. IOP Publishing, 052031.

[24] I. Rhee, L. Xu, S. Ha, A. Zimmermann, L. Eggert, and R. Scheffenegger. 2018. CUBIC for Fast Long-Distance Networks. RFC 8312 (Informational). https://doi.org/10.17487/RFC8312

[25] Klaus Schneider, Cheng Yi, Beichuan Zhang, and Lixia Zhang. 2016. A practical congestion control scheme for named data networking. In *Proceedings of the 3rd ACM Conference on Information-Centric Networking*. 21–30.

[26] Pranab Kumar Sen. 1968. Estimates of the regression coefficient based on Kendall's tau. *Journal of the American statistical association* 63, 324 (1968), 1379–1389.

[27] Susmit Shannigrahi, Chengyu Fan, and Christos Papadopoulos. 2018. Named data networking strategies for improving large scientific data transfers. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 1–6.

[28] Susmit Shannigrahi, Christos Papadopoulos, Edmund Yeh, Harvey Newman, Artur Jerzy Barczyk, Ran Liu, Alex Sim, Azher Mughal, Inder Monga, Jean-Roch Vlimant, et al. 2015. Named data networking in climate research and HEP applications. In *Journal of Physics: Conference Series*, Vol. 664. IOP Publishing, 052033.

[29] Junxiao Shi. 2017. *Named Data Networking in Local Area Networks*. Ph.D. Dissertation. The University of Arizona. http://hdl.handle.net/10150/625652

[30] Won So, Ashok Narayanan, and David Oran. 2013. Named data networking on a router: Fast and DoS-resistant forwarding with hash tables. In *Architectures for Networking and Communications Systems*. IEEE, 215–225.

[31] Junji Takemasa, Yuki Koizumi, and Toru Hasegawa. 2017. Toward an ideal NDN router on a commercial off-the-shelf computer. In *Proceedings of the 4th ACM Conference on Information-Centric Networking*. 43–53.

[32] H Thiel. 1950. A rank-invariant method of linear and polynomial regression analysis, Part 3. In *Proceedings of Koninalijke Nederlandse Akademie van Weinenschatpen A*, Vol. 53. 1397–1412.

[33] Matteo Varvello, Diego Perino, and Leonardo Linguaglossa. 2013. On the design and implementation of a wire-speed pending interest table. In *2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 369–374.

[34] Yi Wang, Keqiang He, Huichen Dai, Wei Meng, Junchen Jiang, Bin Liu, and Yan Chen. 2012. Scalable name lookup in NDN using effective name component encoding. In *2012 IEEE 32nd International Conference on Distributed Computing Systems*. IEEE, 688–697.

[35] Edmund Yeh, Ran Liu, Yuanhao Wu, Volkan Mutlu, Yuezhou Liu, Harvey Newman, Catalin Iordache, Raimondas Sirvinskas, Justas Balcas, Susmit Shannigrahi, Chengyu Fan, and Craig Partridge. 2019. SANDIE: SDN-Assisted NDN for Data Intensive Experiments. In *SC19 Network Research Exhibition*.

[36] Wei You, Bertrand Mathieu, Patrick Truong, Jean-François Peltier, and Gwendal Simon. 2012. Dipit: A distributed bloom-filter based pit table for ccn nodes. In *2012 21st International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 1–7.

[37] Haowei Yuan and Patrick Crowley. 2014. Scalable pending interest table design: From principles to practice. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2049–2057.

[38] Haowei Yuan, Tian Song, and Patrick Crowley. 2012. Scalable NDN forwarding: Concepts, issues and principles. In *2012 21st International Conference on computer communications and networks (ICCCN)*. IEEE, 1–9.

[39] Lixia Zhang, Alexander Afanasyev, Jeffrey Burke, Van Jacobson, KC Claffy, Patrick Crowley, Christos Papadopoulos, Lan Wang, and Beichuan Zhang. 2014. Named Data Networking. *ACM SIGCOMM Computer Communication Review* 44, 3 (2014), 66–73.

Shi, Junxiao; Pesavento, Davide; Benmohamed, Lotfi. "NDN-DPDK: NDN Forwarding at 100 Gbps on Commodity Hardware." Paper presented at 7th ACM Conference on Information-Centric Networking (ICN 2020), Montreal, CA. September 29, 2020 - October 01, 2020.

A Document-based View of the Risk Management Framework

Joshua Lubell, National Institute of Standards and Technology <joshua.lubell@nist.gov>

Abstract

Cybersecurity professionals know the Risk Management Framework (RMF) as a rigorous yet flexible process for managing security risk. But the RMF lacks a document focus, even though much of the process requires authoring, reviewing, revising, and accessing plans and reports. It is possible to build such a focus by looking more closely at these documents, starting with the *System Security Plan* and the roles of key participants responsible for it. Such a document- and role-centric view of the RMF process can lead the way toward more efficient and less error-prone security assurance.

Table of Contents

1. Introduction 1
2. RMF-derived System Security Plan Workflows 3
    2.1. RMF Steps and High-level Workflow 3
    2.2. System Security Plan and Sample Document Model 4
    2.3. Key Participants 6
    2.4. Step-level Workflows 6
    2.5. A “Publishing” Perspective 14
3. Related Work 15
    3.1. Incentive-focused Document Workflow Analysis 16
    3.2. DITA 16
    3.3. Cybersecurity System Risk Indicator (CSRI) 17
4. Conclusion 17
A. OSCAL Representation of AU-1 Subset 18
References 18

1. Introduction

The National Institute of Technology's (NIST's) Risk Management Framework (RMF) [JTFTI2018] defines a rigorous yet flexible process for managing security risk. Application of the RMF process provides the evidence needed to justify assurance that an information system is operating within acceptable risk tolerance. The United States Government's Federal Information Security Modernization Act (FISMA) mandates RMF use for federal agencies and their contractors, yet the RMF process is sufficiently flexible to be used by all sorts of organizations. For example, Graves et al. show how an additive manufacturing service provider can use the RMF to assess system security risk [Graves] .

The RMF lists the roles and responsibilities (summarized in Section 2.3) of those primarily responsible for managing an organization's risk. Yet it is up to the organization whether multiple people fill a single role, whether a single person fills multiple roles, or whether a role is outsourced. Such a decision is based on multiple factors, including size of the organization, its appetite for risk, budget constraints, regulatory requirements, and the consequences of a loss of confidentiality, integrity, or availability (CIA) of the information its systems store, process, or transmit. One extreme might be a large government agency

Lubell, Joshua. "A Document-based View of the Risk Management Framework." Paper presented at Balisage: The Markup Conference, Washington, DC, US. July 27, 2020 - July 31, 2020.

or corporation with an entire department dedicated to information security risk management, with teams responsible for each role. The opposite extreme might be a family-run small business where all operational roles are filled by a single technology-savvy employee or outside service provider, and executive roles are filled by an owner, who consults with outside experts as needed. Regardless of whether the organization is large or small, public or private sector, or deals with information where a loss of CIA would be catastrophic to many people or few people, the RMF process establishes who is responsible and accountable for information security assurance.

Executing the RMF process requires preparation, modification, and review of a variety of documents. Although the RMF is structured and precise in describing its process, it describes these documents in an unstructured and disjointed manner. The documents are referenced in the context of the risk management tasks that involve them and of the people who are responsible for preparing, modifying, and reviewing them. But the RMF provides no schema or declarative markup vocabulary for these documents. It is therefore up to organizations using the RMF to provide document authoring guidance. Some organizations use spreadsheet or word processor document templates. Others use software tools such as the United States government's Cyber Security Evaluation Tool [CSET], which uses a questionnaire-driven approach to produce a Security Assessment Report document.

However, these templates and tools are one-off efforts that do not interoperate. A document created using one of them cannot be easily imported into another. Also, a template or tool designed for one organization's document may be unusable for another organization. And while a template provides guidance on formatting and structure, it has only limited ability to validate a document's completeness or consistency. As a result, RMF users lack the good authoring and content management tools they need to efficiently edit, navigate, share, and evaluate the very documents that are central to system security assurance.

As an initial step toward remedying the lack of tool support, this paper demonstrates a way to derive from the RMF a more document-centric view of risk management. This envisioned document-based view supplements the existing RMF guidance by providing an alternative way of looking at security assurance that is less process-focused and more oriented toward publishing, i.e., document authoring and management. Such a document focus can facilitate the development of schemas, tools for authors, reviewers, and approvers, and content management systems to better support distribution and archiving of risk management information. Beneficiaries would be the people filling the risk management roles listed in Section 2.3. This paper does not attempt to define machine-readable document models for the RMF. That goal, which is too ambitious for a single paper (or single person for that matter) is a long-term goal of NIST's Open Security Controls Assessment Language (OSCAL) project [OSCAL] [Piez19]. However, the methodology this paper demonstrates could be useful to projects such as OSCAL as a means for ensuring that such document models are faithful to the RMF process-centric guidance.

But why choose NIST's RMF for this investigation when there are other frameworks for system security assurance such as the ISO 27000 [ISO27000] and ISO/IEC 15408 [ISO15408] families of International Standards? This paper chooses the RMF as its basis because the RMF is widely used within the federal government and voluntarily used outside the federal government. And the RMF does not exist in a vacuum. The RMF's system life cycle process and use of systems engineering terminology are derived from ISO/IEC15288 [ISO15288], a systems engineering standard covering processes and life cycle stages [Zemrowski]. Also, some RMF tasks can be executed using NIST's “Framework for Improving Critical Infrastructure Cybersecurity” [NIST], a voluntary risk-based management approach used widely in both private and public sectors internationally. Additionally, the RMF process is compatible with any of the standardized catalogs of security controls. A security control is a safeguard or countermeasure that protects the confidentiality, integrity, and availability of a system and its information. Although the RMF refers users to the NIST Special Publication (SP) 800-53 [JTFTI2013] guidance for selection and tailoring of security controls, users not subject to FISMA are free to substitute security controls from ISO 27001 [ISO27001], ISO/IEC 15408, or other sources.

This paper proceeds as follows. The next section provides a high-level introduction to the RMF process and illustrates a document-focused view of RMF focusing specifically on the *System Security Plan* —

representative of the various documents involved in the RMF process — and its associated workflows. The third section discusses other relevant research and development contributions. The last section concludes the paper.

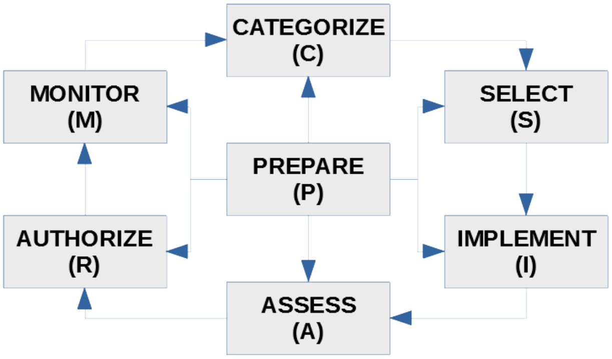
2. RMF-derived System Security Plan Workflows

The RMF defines a process that organizations can use to manage risk at both the enterprise-wide and system levels. The process is iterative and adaptive in response to new threats or changes to an organization's mission or business functions. The RMF process has a series of seven steps: *Prepare*, *Categorize*, *Select*, *Implement*, *Assess*, *Authorize*, and *Monitor*. Each step consists of multiple tasks. Each task has a set of potential inputs, some of which are expected outputs from other tasks. A task output that serves as an input to another task is often a document, although the RMF does not recommend specific document formats or schemas. These documents and the choreography between parties responsible for producing, reviewing, modifying, and consuming them constitute a set of workflows.

2.1. RMF Steps and High-level Workflow

Figure 1 illustrates how the seven RMF steps relate to one another and form a high-level workflow, with the caveat that steps are not always required to be followed in order. Also, although the connectors between steps are unidirectional, it is sometimes necessary to go backwards. For example, as discussed in Section 2.4.5, an assessment may trigger a re-implementation of security controls. The *Prepare* step's tasks lay the groundwork necessary to carry out the other RMF steps accurately and efficiently. Because the RMF process is iterative and adaptive, a task outcome from one of the six other steps may require revisiting one or more *Prepare* tasks.

Figure 1.



Risk Management Framework steps [JTFTI2018].

The six other RMF steps each serve the following purposes:

Categorize Describes the system and classifies the impact of loss of confidentiality, integrity, and availability of the information it stores, processes, and transmits.

Select Chooses an initial set of controls for the system, tailoring them as needed to reduce the risk to an acceptable level.

Implement Implements the controls and describes their deployment.

Assess Determines whether the controls are implemented correctly and are fit for purpose.

Authorize Determines whether the system is safe to operate or use based on acceptability of risk to operations, assets, and people. Authorization in this context is not to be confused with authorizing a user/process/device access to a system's resources. The former is a risk management function. The latter is a system function that should be subject to security controls enforcing identity management and access control policies.

Monitor Continuously monitors the system and associated controls to assess control effectiveness, report changes to the system and its environment, assess risk and impact, and report security posture.

The RMF is process-oriented and task-oriented. It is not document-oriented, although many task inputs and outputs are documents. As such, RMF guidance is organized by steps and their tasks rather than organized by document. Therefore, RMF specifies document workflows implicitly rather than explicitly. Extracting an implicit document workflow involves analyzing task descriptions that mention the document or a subset of the document as a potential input or expected output.

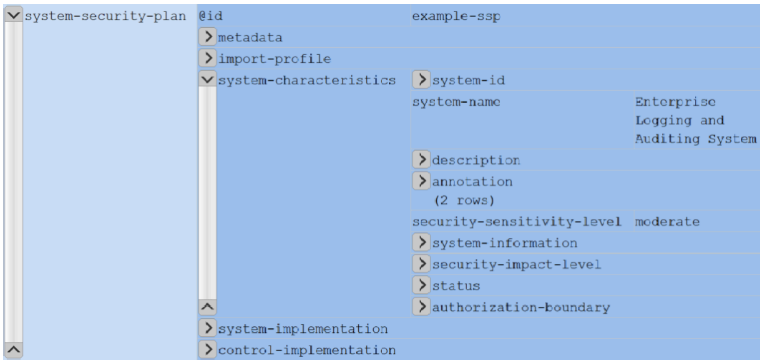
2.2. System Security Plan and Sample Document Model

This paper studies the *System Security Plan (SSP)* and its associated workflow based on the RMF tasks most involved with producing it. The SSP is only one of several documents integral to the RMF process. However, SSPs cover a significant subset of RMF tasks and involve all seven RMF steps. The SSP documents the security requirements for an information system and describes the security controls in place or planned for meeting those requirements [OMB]. The “system” in an SSP may be an individual workstation or laptop, a server, or a networked device. Networked devices can include operational technology such as “Internet of Things” devices, 3D printers, digitally controlled machines, industrial switches, and programmable logic controllers. Alternatively, a system may be a logically grouped collection of computers and/or devices. Thus, a single SSP can apply to more than one computer or device.

The RMF uses prose text to describe the information required for an SSP but provides no machine-readable schema for automated syntactic validation of whether an SSP is complete. This paper uses the current draft form of the OSCAL architecture, which includes an SSP document model, as an alternative approach for supporting the RMF SSP documentation objectives and enabling automated validation.

The OSCAL GitHub repository includes, as an example, an Extensible Markup Language (XML) document valid with respect to this SSP model. This simple example represents the SSP of a partially-implemented system that provides enterprise logging and log auditing capabilities and uses controls from the SP 800-53 catalog. This paper's analysis of step-level workflows (Section 2.4) uses this example as a means of showing how individual tasks within each RMF step affect the SSP contents. Figure 2 shows a high-level view of the example using a spreadsheet-like grid.

Figure 2.



```
system-security-plan  @id                     example-ssp
                      metadata
                      import-profile
                      system-characteristics  system-id
                                              system-name                  Enterprise
                                                                           Logging and
                                                                           Auditing System
                                              description
                                              annotation
                                                (2 rows)
                                              security-sensitivity-level  moderate
                                              system-information
                                              security-impact-level
                                              status
                                              authorization-boundary
                      system-implementation
                      control-implementation
```

High-level OSCAL SSP document model example.

Some of the major SSP document model elements are defined as follows:

`import-profile`
References a pre-defined set of security requirements, for example a baseline from SP 800-53.

`security-sensitivity-level`
Indicates the system's overall information sensitivity categorization. May be “low”, “moderate”, or “high”.

`system-information`
Describes all the information types the system stores, processes, or transmits, for example, historical logging and auditing information. Each description specifies the impact of a loss of confidentiality, integrity, and availability of the information type.

`security-impact-level`
Specifies target levels of confidentiality, integrity, and availability for the system.

`authorization-boundary`
Establishes a system's scope of protection. Authorization boundary determination is the act of specifying what the organization is directly responsible for protecting.

`system-implementation`
Specifies the types of users who interact with the system and their roles, the system's component parts, services (e.g., ports and protocols), interconnections, and inventory items.

`control-implementation`
Describes how the system implements a set of controls. For each control implemented, specifies the control's description and set of implemented requirements satisfied.

`component`
Defines a part of an implemented system. A `component` can be hardware, software, or a service, policy, process, or procedure. Although not shown in Figure 2, components are critical model elements in that they enable the expression of relationships between implemented controls, hardware and software assets, and policies and business processes.

2.3. Key Participants

RMF Appendix D [JTFTI2018] lists the roles and responsibilities of key participants in the risk management process. These include the following key participants primarily responsible for RMF tasks impacting the SSP.

Information Security Officer	Oversees security responsibilities at the organizational level.
Authorizing Official	Assumes accountability for operating a system. May empower a designated representative to carry out many of the activities related to the execution of the RMF. However, only the Authorizing Official can determine whether the risk from the operation or use of the system is acceptable.
System Owner	Buys, develops, integrates, modifies, operates, maintains, and disposes of a system. Responsible for creating and maintaining the SSP.
Information Owner or Steward	Establishing the rules for appropriate use and protection of a specified type of information. A system may contain information from multiple information owners or stewards.
Common Control Provider	Implements, assesses, and monitors common controls. Common controls can be inherited by multiple systems, thus reducing the complexity and protection costs of an organization's IT infrastructure. For example, hardware token-based authentication, implemented enterprise-wide using Personal Identity Verification cards, is an example of a common control.
Security Architect	Ensures that the enterprise architecture addresses stakeholder protection needs and the corresponding system requirements necessary to protect organizational missions and business functions.
Control Assessor	Evaluates implemented controls to determine their effectiveness.

For a small organization, one person might fulfill multiple roles. Conversely, a large organization could have multiple people filling a single role, for example, a team of Control Assessors. Also, some roles can be outsourced to third parties, such as Common Control Provider, Control Assessor, and System Owner (as in the case of a cloud-based service). Other roles, such as Authorizing Official, cannot be outsourced as they require executive-level accountability.

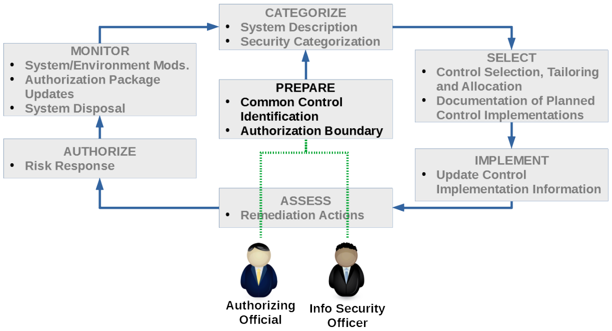
2.4. Step-level Workflows

The following subsections describe the SSP publishing workflow centered around each RMF step. A figure highlighting the RMF step illustrates each workflow. The highlighted RMF step has a bulleted list of the subset of RMF tasks for that step directly pertaining to the SSP. This bulleted list does not exist explicitly in SP 800-37 but is derived from the subset of expected outputs that involve SSP content authoring, SSP content review, or determination of SSP approval. The figure also shows key participants responsible for these tasks with dashed connectors pointing either to the RMF step or to a directional arrow leading to or from the step. A second figure indicates which portions of the SSP document model described in Section 2.2 get populated or modified in each workflow. The *Categorize* and *Select* publishing workflows (Section 2.4.2 and Section 2.4.3) also include XML code illustrating how the `component` element enables traceability between control implementations, assets, business processes, and policies.

2.4.1. Prepare

Two *Prepare* tasks that result in modification to the SSP are “Common Control Identification” and “Authorization Boundary” determination. As shown in Figure 3, the Information Security Officer and Authorizing Official add content to the SSP. The Information Security Officer identifies common controls at the organizational level and documents their planned implementation. The Authorizing Official determines and documents a system's authorization boundary, using the system design documentation as input.

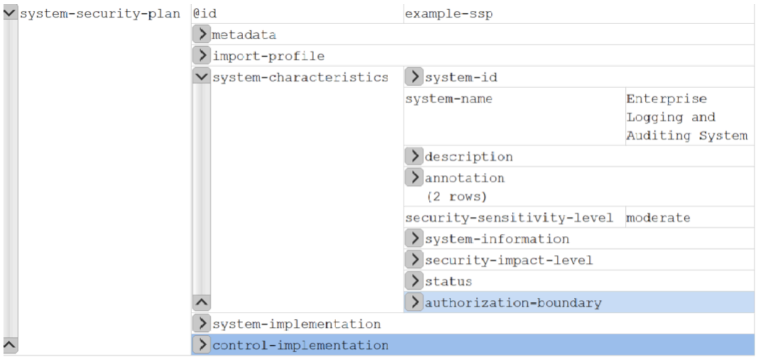
Figure 3.


CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
PREPARE
• Common Control Identification
• Authorization Boundary
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions
Authorizing Official
Info Security Officer

SSP interactions during *Prepare* step.

Figure 4 indicates the portions of the SSP document where new information is added, namely the `authorization-boundary` element and the portion of the `control-implementation` element pertaining to common control documentation. System design documentation comprises much of the input to the “Authorization Boundary” task. The RMF provides no guidance on this documentation's contents or structure but suggests that it includes a mix of prose and diagrams defining and identifying system elements, their interactions and the environment in which the system operates. Thus, `authorization-boundary` might reference a diagram showing an authorization boundary including the server and logging software with an environment of operation including client devices or services that write to or read from the log. `control-implementation` specifies any common controls inherited by the logging and auditing system. For example, if the system is located in a facility with physical access controls, `control-implementation` would include the physical access controls. The *Select* step-level workflow (Section 2.4.3) includes an example of `control-implementation` XML markup that does not specify a common control, but rather is specific to the logging and auditing system.

Figure 4.



```
system-security-plan  @id                      example-ssp
                      metadata
                      import-profile
                      system-characteristics   system-id
                                               system-name                  Enterprise
                                                                            Logging and
                                                                            Auditing System
                                               description
                                               annotation
                                                 (2 rows)
                                               security-sensitivity-level   moderate
                                               system-information
                                               security-impact-level
                                               status
                                               authorization-boundary
                      system-implementation
                      control-implementation
```

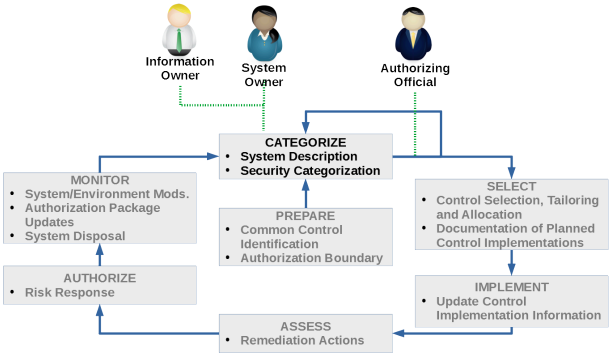
Affected portions of SSP document.

2.4.2. Categorize

Two *Categorize* tasks that result in modification to the SSP are “System Description” and “Security Categorization”. As shown in Figure 5, the System Owner, Information Owner or Steward, and Authorizing Official bear primary responsibility for tasks impacting the SSP contents. The System Owner is primarily responsible for the System Description task and, together with the Information Owner or Steward, bears joint responsibility for the Security Categorization task. The Authorizing Official reviews the security categorization results and decides whether to approve the categorization. If approved, the RMF process proceeds with the *Select* step. Otherwise, the categorization process must be repeated.

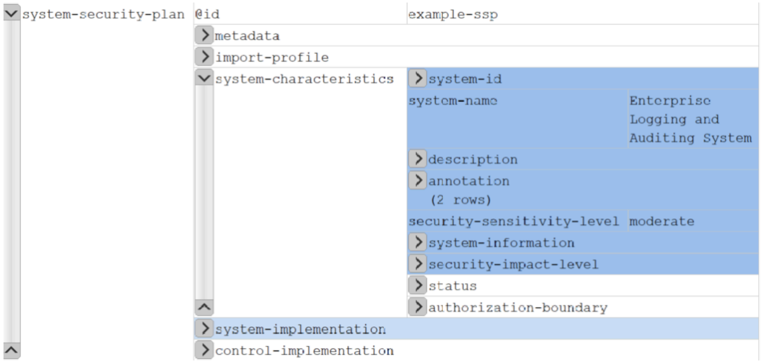
As shown in Figure 6, the *Categorize* step populates a substantial portion of the SSP. The two tasks together create content for the `system-implementation` element and most of the `system-characteristics` element content except for its `status` and `authorization-boundary` sub-elements. `system-information`, a child element of `system-characteristics` not shown in Figure 6, lists a single information type: “System and Network Monitoring”. The specified impact of a loss of confidentiality, integrity, and availability is obtained by referencing NIST's “Appendices to Guide for Mapping Types of Information and Information Systems to Security Categories” [Stine], which provides a catalog of information types with suggested impacts.

Figure 5.


Information Owner
System Owner
Authorizing Official
CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
PREPARE
• Common Control Identification
• Authorization Boundary
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions

SSP interactions during *Categorize* step.

Figure 6.



system-security-plan | @id | example-ssp
metadata
import-profile
system-characteristics
system-id
system-name | Enterprise Logging and Auditing System
description
annotation
(2 rows)
security-sensitivity-level | moderate
system-information
security-impact-level
status
authorization-boundary
system-implementation
control-implementation

Affected portions of SSP document.

`system-implementation` contains many descendant elements representing the system's components, assets, and roles responsible for these entities. The OSCAL SSP example defines several logging and auditing system components. These include the server software, the enterprise logging, monitoring, and alerting policy, the systems integration and inventory management processes, and the enterprise's configuration management guidance. In the interest of brevity, this paper only shows XML markup and content pertaining to the logging, monitoring, and alerting policy. The following XML defines the policy component, assigning it an identifier and assigning responsibility for maintaining the policy to the legal department.

```
<component id="component-logging-policy" component-type="policy">
    <prop name="version">2.1</prop>
    <prop name="last-modified-date">20181015</prop>
    <status state="operational"/>
    <responsible-role role-id="maintainer">
        <party-id>legal-department</party-id>
    </responsible-role>
</component>
```

The XML below defines the logging server asset as part of the system's asset inventory. The logging server is assigned an administrator and owner. `implemented-component` references a component implemented in a given system inventory item. Thus, the logging server implements both the server software and policy components.

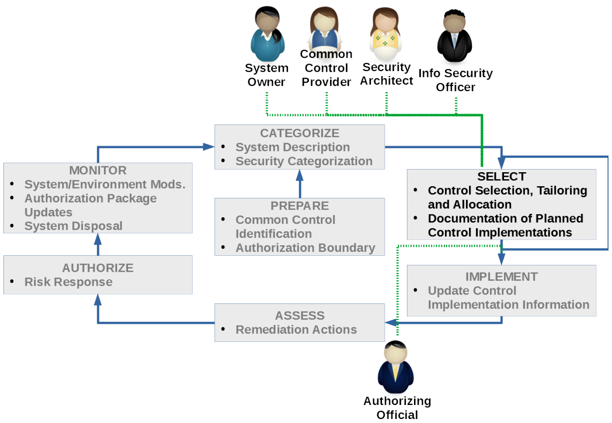
```
<inventory-item id="inventory-logging-server"
                asset-id="asset-id-logging-server">
    <responsible-party role-id="asset-administrator">
        <party-id>enterprise-asset-administrators</party-id>
    </responsible-party>
    <responsible-party role-id="asset-owner">
        <party-id>enterprise-asset-owners</party-id>
    </responsible-party>
    <implemented-component component-id="component-logging-server"
                           use="runs-software"/>
    <implemented-component component-id="component-logging-policy"
                           use="enforces-policy"/>
</inventory-item>
```

Lubell, Joshua. "A Document-based View of the Risk Management Framework." Paper presented at Balisage: The Markup Conference, Washington, DC, US. July 27, 2020 - July 31, 2020.

2.4.3. Select

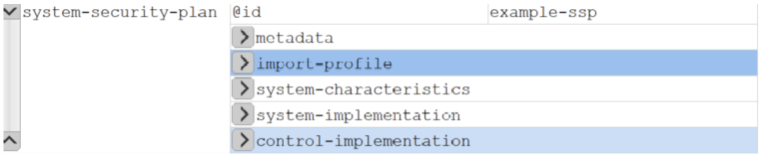
The *Select* tasks that result in modification to the SSP (shown in Figure 7) are “Control Selection, Tailoring, and Allocation” and “Documentation of Planned Control Implementations”. The System Owner and Common Control Provider jointly bear primary responsibility for selecting and tailoring the system's security controls. The Security Architect and Information Security Officer are jointly responsible for control allocation. Allocation entails determining whether controls shall be system-specific, hybrid, or common, and then assigning the controls to the system elements (i.e., machine or environment in which the system operates) responsible for providing a security capability. The System Owner and Common Control Provider are responsible for documenting the controls and plans for their implementation in the SSP, as represented in the `control-implementation` element in Figure 8. `import-profile` contains a reference to common controls (such as the physical access controls mentioned in Section 2.4.1) identified during the *Prepare* step.

Figure 7.


System Owner
Common Control Provider
Security Architect
Info Security Officer
CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
PREPARE
• Common Control Identification
• Authorization Boundary
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions
Authorizing Official

SSP interactions during *Select* step.

Figure 8.



```
system-security-plan   @id                    example-ssp
                       metadata
                       import-profile
                       system-characteristics
                       system-implementation
                       control-implementation
```

Affected portions of SSP document.

`control-implementation` is updated to include additional controls needed to supplement the inherited common controls. The System Owner selects SP 800-53 control AU-1 (Audit and Accountability Policy and Procedures), whose control statement is shown in Figure 9. The italicized text inside brackets represents parameters. Selecting a SP 800-53 control requires inserting values for these parameters. Appendix A provides the AU-2 control statement as represented in OSCAL's SP 800-53 catalog model. As Appendix A shows, the OSCAL catalog model uses a `part` element to represent each of the nested list items in Figure 9. Each `part` has an `@id` attribute. Each of the three parameters is represented by a `param` element, also with an `@id` attribute.

Figure 9.

> The organization:
>
> a. Develops, documents, and disseminates to [*organization-defined personnel or roles*]:
>
> i. An audit and accountability policy that addresses purpose, scope, roles, responsibilities, management commitment, coordination among organizational entities, and compliance; and
>
> ii. Procedures to facilitate the implementation of the audit and accountability policy and associated audit and accountability controls; and
>
> b. Reviews and updates the current:
>
> i. Audit and accountability policy [*organization-defined frequency*]; and
>
> ii. Audit and accountability procedures [*organization-defined frequency*].

Security control AU-1: Audit and Accountability Policy and Procedures [JTFTI2013].

The SSP documents the selection and planned implementation of AU-1 by associating each part of the control statement with the component that implements the part, assigning parameter values as needed. The following XML represents the planned implementation of list item "b.", sub-list item "i." from Figure 9 ("The organization … Reviews and updates the current … Audit and accountability policy [*organization-defined frequency*]").

```
<control-implementation>
    <implemented-requirement control-id="au-1">
      ...
      <statement statement-id="au-1_smt.b.1">
          <by-component component-id="component-logging-policy">
              <set-parameter param-id="au-1_prm_2">
                   <value>annually, and other times as necessary in
response to regulatory or organizational changes</value>
              </set-parameter>
          </by-component>
      </statement>
      ...
   </implemented-requirement>
</control-implementation>
```

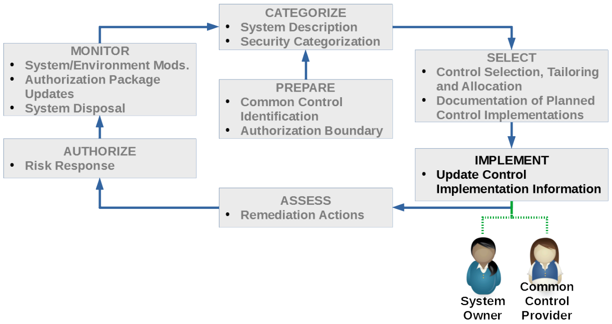
`by-component` references the component implementing this portion of the AU-1 statement. The *organization-defined frequency* parameter is assigned the value "annually, and other times as necessary in response to regulatory or organizational changes".

Once the controls and their planned implementation are documented, the Authorizing Official reviews the SSP to determine if it is complete, consistent, and satisfies the system's security requirements. If approved, the RMF process proceeds with the *Implement* step. Otherwise, the *Select* step must be repeated.

2.4.4. Implement

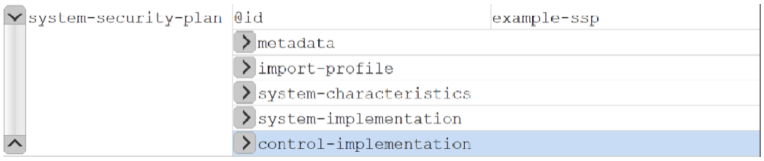
As the final task culminating the *Implement* step, the System Owner and Common Control Provider update the SSP's control implementation information (shown in Figure 10) to reflect any differences between the actual control implementations and the planned implementation documentation produced in the *Select* step. Figure 11 shows that the update results in a modification to the SSP's `control-implementation` element as needed to reflect the actual implementation.

Figure 10.


CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
PREPARE
• Common Control Identification
• Authorization Boundary
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions
System Owner
Common Control Provider

SSP interactions during *Implement* step.

Figure 11.



```
system-security-plan   @id                    example-ssp
                       metadata
                       import-profile
                       system-characteristics
                       system-implementation
                       control-implementation
```

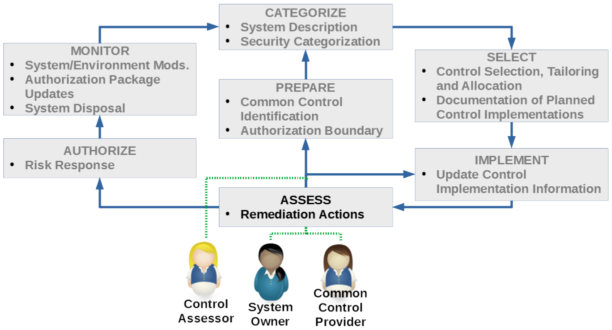
Affected portions of SSP document.

2.4.5. Assess

During the *Assess* step (Figure 12), the System Owner and Common Control Provider update the SSP to reflect the state of the controls after the Control Assessor's initial assessment and any changes made in response to recommended “Remediation Actions”. If the assessment indicates controls were not properly implemented, the *Implement* step needs to be revisited. The assessment findings could also possibly trigger an update to a system-level or organization-wide risk assessment, requiring a return to the *Prepare* step.

As was the case for the *Implement* step (Figure 11), the *Assess* step's effect on the SSP document is limited to modification of the `control-implementation` element.

Figure 12.

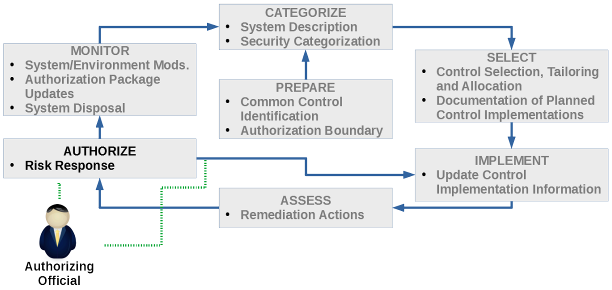

CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
PREPARE
• Common Control Identification
• Authorization Boundary
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions
Control Assessor
System Owner
Common Control Provider

SSP interactions during *Assess* step.

2.4.6. Authorize

Following an analysis of risk from operating or using the system, the Authorizing Official determines whether the response to the risk is acceptable. The "Risk Response" is based on a review of the System Owner's and Common Control Provider's modification to the SSP (Figure 11), and of related documents such as assessment reports and Plan of Action and Milestones (defined in Section 3.3) for addressing any SSP deficiencies. These documents together comprise the Authorization Package. A determination that the Risk Response is unacceptable results in a return to the *Implement* step, as shown in Figure 13.

Figure 13.

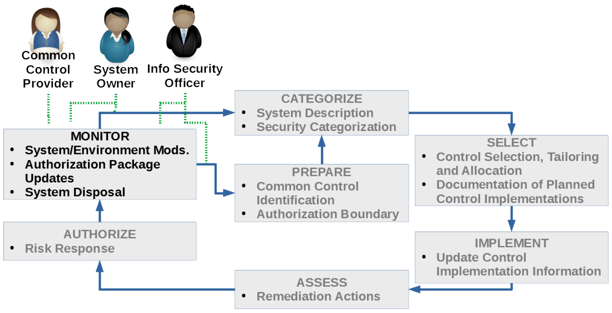

CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
PREPARE
• Common Control Identification
• Authorization Boundary
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions
Authorizing Official

SSP interactions during *Authorize* step.

2.4.7. Monitor

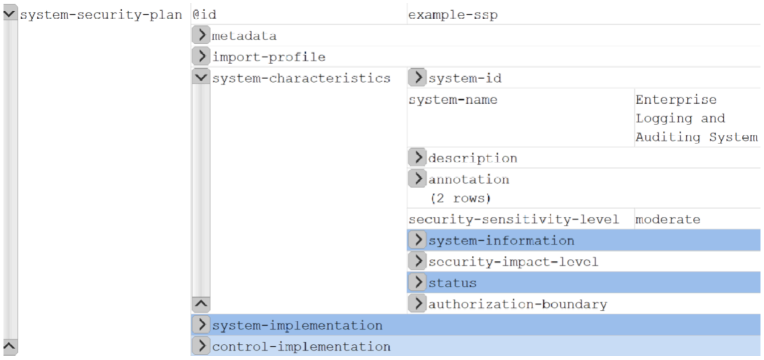
There are three tasks in the *Monitor* step that can trigger updates to the SSP, as shown in Figure 14. They are "System and Environment Changes", "Authorization Package Updates", and "System Disposal". Examples of system and environmental changes include machine elements such as hardware or software upgrades, human elements such as staff turnover, and environmental or physical elements such as physical access controls or relocation of the facility. These changes result in the System Owner or Common Control Provider updating `system-information` and `system-implementation` (as shown in Figure 15) and a return to the *Categorize* step. The *Prepare* step must be revisited as well if determined necessary by the Information Security Officer.

Figure 14.


Common Control Provider
System Owner
Info Security Officer
CATEGORIZE
• System Description
• Security Categorization
MONITOR
• System/Environment Mods.
• Authorization Package Updates
• System Disposal
SELECT
• Control Selection, Tailoring and Allocation
• Documentation of Planned Control Implementations
PREPARE
• Common Control Identification
• Authorization Boundary
AUTHORIZE
• Risk Response
IMPLEMENT
• Update Control Implementation Information
ASSESS
• Remediation Actions

SSP interactions during *Monitor* step.

Figure 15.



```
system-security-plan  @id                          example-ssp
                      metadata
                      import-profile
                      system-characteristics       system-id
                                                   system-name                  Enterprise
                                                                                Logging and
                                                                                Auditing System
                                                   description
                                                   annotation
                                                    (2 rows)
                                                   security-sensitivity-level   moderate
                                                   system-information
                                                   security-impact-level
                                                   status
                                                   authorization-boundary
                      system-implementation
                      control-implementation
```

Affected portions of SSP document.

To achieve timely risk management, the System Owner and Common Control Provider update the SSP included in the Authorization Package in response to continuous monitoring results. These “Authorization Package Updates”, like “System and Environment Changes”, affect the RMF workflow and SSP contents as shown in Figure 14 and Figure 15, respectively.

System Disposal — removal of a system from operation — requires multiple actions (e.g., media sanitization, configuration management, record retention) for which the System Owner bears primary responsibility. These actions result in updating the `status` element in the SSP as shown in Figure 15.

2.5. A “Publishing” Perspective

The previous subsection analyzed the RMF workflows from an SSP document perspective. This subsection looks at the analysis results from the viewpoint of three activities central to SSP document “publishing”: authoring, reviewing, and updating. *Authoring* is the creation of new SSP document content. *Reviewing* is the evaluation of SSP content to determine whether it meets a set of criteria. *Updating* is the modification of existing SSP content in response to reviewer feedback, new implementation experience, or an environmental change. Table 1 summarizes the results of this publishing-focused analysis.

SSP authoring takes place mainly during *Categorize* and *Select*. As shown in Section 2.4.2 and Section 2.4.3, most of the SSP content is created during these RMF steps, by the System Owner and — to the extent that the system leverages common controls — Common Control Provider. Some authoring also takes place during *Prepare* when the Information Security Officer identifies common controls and documents their planned implementation. The Authorizing Official documents the authorization boundary during *Prepare*, but this consists mostly of references to concepts from other documents such as system design documents and diagrams.

SSP reviewing occurs during most RMF steps, but especially during *Assess* and *Authorize* where reviewing is the central purpose of the step. The Authorizing Official is the chief reviewer of the RMF process and has the greatest authority and accountability. The Control Assessor and Information Security Officer have important reviewer roles as well. These three reviewers have differing and complementary qualifications. The Authorizing Official is a senior executive or business owner with an intimate understanding of organizational mission, budget constraints, and security and privacy risks. The Control Assessor is an information security expert with the detailed knowledge necessary to evaluate effectiveness of control implementations. The Information Security Officer has expertise spanning both security assurance and implementation and offers both an organization and systems perspective.

SSP updating occurs during *Implement*, *Assess*, and *Monitor*, with the System Owner and possibly the Common Control Provider bearing primary responsibility. Given that the System Owner and Common

Control Provider also have outsized roles in SSP authoring, it follows that their requirements should be given top priority when developing SSP document schemas and editing tools.

Table 1.

SSP authoring, reviewing, and updating throughout the RMF process.

Activity	RMF Step	Primary Responsibility
Authoring	*Prepare*	Information Security Officer, Authorizing Official
	Categorize	System Owner, Common Control Provider, Information Owner or Steward
	Select	System Owner, Common Control Provider, Security Architect, Information Security Officer
Reviewing	*Categorize*, *Select*	Authorizing Official
	Assess	Control Assessor
	Authorize	Authorizing Official
	Monitor	Information Security Officer
Updating	*Implement*, *Assess*, *Monitor*	System Owner, Common Control Provider

This paper claimed in Section 1 that a document-centric analysis can lead to improved schemas and tools for the key participants in the RMF process. The analysis in Section 2.4 and subsequent analysis in this subsection support that claim by providing these takeaways:

- Authoring schemas and editing tools should prioritize the needs of the System Owner and Common Control Provider. People fulfilling these roles may have technical knowledge about systems and/or controls for which they are responsible. They are less likely, however, to have risk management or security assurance expertise. Because these roles produce and maintain much of the SSP content, making their authoring and update tasks easier and less error-prone should result in better cybersecurity and cost savings to their organizations.

- Solutions that automate review/update workflows have the challenges of accommodating not only reviewers, but also authors performing updates. Additionally, these solutions should be tailored to the differing needs and areas of expertise of the Authorizing Official, Control Assessor, Information Security Officer reviewer roles.

- Since no single tool can fit the requirements of all key participants in the RMF process, interoperability standards are needed to support an ecosystem of tools. The proposed and evolving OSCAL document formats could meet this need. Another candidate is the Darwin Information Typing Architecture, discussed in Section 3.2.

3. Related Work

Each of the following research and development efforts offer insights or technologies relevant to this paper's document-based analysis of the RMF. The first offers lessons learned from a publishing discipline outside the realm of security assurance. The second is a standards architecture that supports publishing workflows and can also enable specialized data models for controls, catalogs, and profiles. The third is a cyber risk measurement technique whose utility would be enhanced by the document-focused view of the RMF this paper advocates.

3.1. Incentive-focused Document Workflow Analysis

A publishing workflow's success depends upon the actors involved being incentivized to work toward a common goal. In the case of RMF-based security assurance, there are two common scenarios. For United States federal agencies and their contractors, the System Owner, Information Security Officer, Authorizing Official, and other participants are all incentivized by FISMA to adhere to the RMF process. Non-federal companies and organizations have no FISMA requirement. However, the RMF's system life cycle approach ensures that security plans and implementations are mission-based and tied to business requirements. Therefore, even without a common incentive such as FISMA compliance, good systems engineering practices can result in mutually reinforcing incentives for those responsible for SSP development and development of other RMF document outputs.

Piez [Piez20] studies workflow participant incentives in research journal publishing enterprises: an industry quite distinct from security assurance. There are some parallels in terms of roles (Author versus System Owner, Peer Reviewer versus Control Assessor, Journal Editor versus Authorizing Official) and process steps (*Review* versus *Assess*, *Accept/Reject* versus *Authorize*). However, there are significant differences in incentives. For example, security assurance strongly encourages identification and implementation of common controls wherever possible. Research articles submitted to journals, on the other hand, are required to have intellectual content that is mostly original. Even survey articles, where previously published work is highlighted, must contain a thorough and substantive analysis of the prior art.

Unlike security assurance, where deployment of declarative document markup technologies such as OSCAL is in its infancy, journal publishing workflows have been XML-enabled for a long time. However, as Piez points out, XML has achieved its greatest successes in post-production processes where authors and reviewers are not involved. Despite decades of XML implementation experience in the journal publishing world, most authors still submit papers in word processor formats, and many use spreadsheets for managing and responding to reviewer comments. Lessons learned regarding what works and what does not from the journal publishing industry could help markup technologists determine the implementation approaches most likely to be successful in automating and streamlining development of SSPs and other security assurance documents. However, implementers of security assurance publishing tools should be aware that incentives among security assurance workflow participants differ from the ecosystem of incentives that exist in the journal publishing world.

3.2. DITA

The Darwin Information Typing Architecture (DITA) [OASIS], a standardized XML-based architecture for authoring, managing, reusing and transforming technical content, could provide the foundation for an interoperable set of tools to meet the needs of SSP authors, updaters, and reviewers. Unlike other declarative markup language frameworks, DITA supports the definition of element types that are *specializations* of other element types [Kimber]. Specializations inherit not only the syntactic structure of their base element type, but also the processing behavior. To understand the benefits of specialization in the context of the RMF process, suppose a DITA element type was defined for control catalogs, with specializations for SP 800-53, ISO 27001, and ISO/IEC 15408. Then a tool developer could leverage the same code implementing the generic catalog element type to support the RMF *Select* step for all three catalogs.

Although OSCAL has a generic catalog schema that can be extended to support other catalogs besides SP 800-53, OSCAL does not support specialization in the general sense. For example, the OSCAL `control-implementation` element's support for parameters (discussed in Section 2.4.3) accommodates SP 800-53 but may not be needed for other catalogs. Therefore, `control-implementation` is *over-specified* for catalogs without control parameters. Conversely, consider a catalog whose controls have properties not pre-defined in OSCAL. `control-implementation`'s content model can be extended using OSCAL's `prop` element to provide name-value pairs for these additional control properties.

However, the OSCAL `control-implementation` element would then be *under-specified* without supplemental documentation for how OSCAL tools should handle the `prop` extensions. DITA's specialization mechanism provides a formal way to avoid over-specification and under-specification, explicitly separating general syntax and processing behavior from specialized syntax and processing behavior.

The DITA Open Toolkit [DITA-OT], an output-producing processor that transforms input authored using DITA-conforming XML vocabularies into other formats, implements an extensible workflow for integrating DITA processing with other tasks. The DITA Open Toolkit could potentially be used for automating the authoring and review of SSPs and other documents in the RMF process. SCAP Composer [Lubell19] [Lubell20], a DITA Open Toolkit plug-in that contributes toward implementing the RMF "System and Environment Changes" task (see Section 2.4.7), provides an example of applying DITA specialization and the DITA Open Toolkit's extensible workflow mechanism to solve a security assurance problem.

3.3. Cybersecurity System Risk Indicator (CSRI)

Wilbanks [Wilbanks] developed a methodology to measure system cybersecurity risk that suggests potential benefits of deploying declarative markup technologies in the security assurance space. Wilbanks's CSRI, used within the United States Department of Education's Federal Student Aid cybersecurity risk management program, employs a set of weighted risk factors to quantify a system's vulnerability to cyber threats. Many of these risk factors are computed by extracting relevant information from documents associated with the RMF process. For example, one highly weighted risk factor involves analyzing a system's Plan of Action and Milestones (POAM), a document created by the System Owner and Common Control Provider based on the Control Assessor's findings and recommendations. The POAM describes planned actions with due dates to correct deficiencies identified during the assessment. The CSRI's POAM risk factor score is based on the number of unmet milestones, their criticality, and aging (how many are overdue).

Because POAMs and other risk management-related documents are typically in word processor or spreadsheet formats, extracting the information needed to compute document-based CSRI risk factors is a time-consuming, manual process. If these documents were tagged in a declarative markup language, they would be more amenable to structured queries and optimized presentation of results. For example, a CSRI dashboard user interface that computes and displays risk factors could be efficiently implemented using low-cost and ubiquitous XML- or HTML5-based technologies.

4. Conclusion

This paper presented a new approach that supports the creation of the System Security Plan, a critical Risk Management Framework output. The new approach, based on an SSP document model defined using declarative markup and on the RMF roles primarily responsible for documenting the SSP, offers an alternative view into security assurance. By emphasizing roles and how they relate to SSP document elements, this alternative view provides a roadmap for implementers of standards and tools for authoring and accessing the information in SSP documents. The same approach used in this paper can be extended to other documents arising from the RMF process such as assessment reports and POAMs. With the research and development results discussed in Section 3, the document and role-centric view of the RMF process can lead the way toward new standards and tools enabling more efficient and less error-prone security assurance.

But, to quote [Piez20], "the platform is not the capability." The best languages, schemas, and tools in the world are supplements — but not substitutes — for the expertise required to assess cyber-risk, the systems engineering skills needed to identify common controls or establish authorization boundaries, and the hardware, software, and people skills needed to implement security controls.

A. OSCAL Representation of AU-1 Subset

The following listing, extracted from the OSCAL SP 800-53 catalog, represents the AU-1 control statement shown in Figure 9.

```
<control class="SP800-53" id="au-1">
    <title>Audit and Accountability Policy and Procedures</title>
    <param id="au-1_prm_1">
        <label>organization-defined personnel or roles</label>
    </param>
    <param id="au-1_prm_2">
        <label>organization-defined frequency</label>
    </param>
    <param id="au-1_prm_3">
        <label>organization-defined frequency</label>
    </param>
    <prop name="label">AU-1</prop>
    <part id="au-1_smt" name="statement">
        <p>The organization:</p>
        <part id="au-1_smt.a" name="item">
            <prop name="label">a.</prop>
            <p>Develops, documents, and disseminates to
<insert param-id="au-1_prm_1"/>:</p>
            <part id="au-1_smt.a.1" name="item">
                <prop name="label">1.</prop>
                <p>An audit and accountability policy that addresses
purpose, scope, roles, responsibilities, management commitment,
coordination among organizational entities, and compliance; and</p>
            </part>
            <part id="au-1_smt.a.2" name="item">
                <prop name="label">2.</prop>
                <p>Procedures to facilitate the implementation of
the audit and accountability policy and associated audit and
accountability controls; and</p></part></part>
        <part id="au-1_smt.b" name="item">
            <prop name="label">b.</prop>
            <p>Reviews and updates the current:</p>
            <part id="au-1_smt.b.1" name="item">
                <prop name="label">1.</prop>
                <p>Audit and accountability policy
<insert param-id="au-1_prm_2"/>; and</p></part>
            <part id="au-1_smt.b.2" name="item">
                <prop name="label">2.</prop>
                <p>Audit and accountability procedures
<insert param-id="au-1_prm_3"/>.</p></part></part></part></control>
```

References

[CSET] “Downloading and Installing CSET.” Accessed May 27, 2020. https://www.us-cert.gov/ics/Downloading-and-Installing-CSET.

[DITA-OT] “DITA Open Toolkit.” Accessed April 18, 2020. https://www.dita-ot.org/.

[Graves] Graves, Lynne M. G., Joshua Lubell, Wayne King, and Mark Yampolskiy. "Characteristic Aspects of Additive Manufacturing Security From Security Awareness Perspectives." IEEE Access 7 (2019). 10.1109/ACCESS.2019.2931738.

[ISO15288] ISO/IEC/IEEE 15288:2015 Systems and software engineering — System life cycle processes.

[ISO15408] ISO/IEC 15408-1:2009 Information technology — Security techniques — Evaluation criteria for IT security — Part 1: Introduction and general model.

[ISO27000] ISO/IEC 27000:2018 Information technology — Security techniques — Information security management systems — Overview and vocabulary.

[ISO27001] ISO/IEC 27001:2013 Information technology — Security techniques — Information security management systems — Requirements.

[JTFTI2013] Joint Task Force Transformation Initiative. "Security and Privacy Controls for Federal Information Systems and Organizations." National Institute of Standards and Technology, April 2013. 10.6028/NIST.SP.800-53r4.

[JTFTI2018] Joint Task Force Transformation Initiative. "Risk Management Framework for Information Systems and Organizations:: A System Life Cycle Approach for Security and Privacy." Gaithersburg, MD: National Institute of Standards and Technology, December 2018. 10.6028/NIST.SP.800-37r2.

[Kimber] Kimber, Eliot. "DITA for Practitioners Volume 1: Architecture and Technology." XMLPress, 2012.

[Lubell19] Lubell, Joshua. "SCAP Composer: A DITA Open Toolkit Plug-in for Packaging Security Content." In *Proceedings of Balisage: The Markup Conference*, Vol. 23. Washington, DC, USA, 2019. 10.4242/BalisageVol23.Lubell01.

[Lubell20] Lubell, Joshua. "SCAP Composer User Guide." National Institute of Standards and Technology, February 28, 2020. 10.6028/NIST.IR.8290.

[NIST] National Institute of Standards and Technology. "Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1." Gaithersburg, MD: National Institute of Standards and Technology, 2018. 10.6028/NIST.CSWP.02122014.

[OMB] Office of Management and Budget. "Managing Information as a Strategic Resource," 2016. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A130/a130revised.pdf.

[OASIS] Organization for the Advancement of Structured Information Standards. "DITA Version 1.3 Specification," 2018. http://docs.oasis-open.org/dita/dita/v1.3/dita-v1.3-part0-overview.html.

[OSCAL] OSCAL: the Open Security Controls Assessment Language. "OSCAL." Accessed April 18, 2020. https://pages.nist.gov/OSCAL/.

[Piez19] Piez, Wendell. "The Open Security Controls Assessment Language (OSCAL): Schema and Metaschema." In *Proceedings of Balisage: The Markup Conference*, Vol. 23. Washington, DC, USA, 2019. 10.4242/BalisageVol23.Piez01.

[Piez20] Piez, Wendell. "Systems security assurance as (micro) publishing: Declarative markup for systems description and assessment." To appear in *Proceedings of Balisage: The Markup Conference*. Washington, DC, USA, 2020.

[Stine] Stine, Kevin, Rich Kissel, William C Barker, Annabelle Lee, and Jim Fahlsing. "Volume II: Appendices to Guide for Mapping Types of Information and Information Systems to Security Categories." National Institute of Standards and Technology, August 2008. 10.6028/NIST.SP.800-60v2r1.

[Wilbanks] Wilbanks, Linda. “What's Your IT Risk Approach?” *IT Professional*, Vol. 20, no. 4 (July 2018): 13–17. 10.1109/MITP.2018.043141663.

[Zemrowski] Zemrowski, Kenneth M. “NIST Bases Flagship Security Engineering Publication on ISO/IEC/IEEE 15288:2015.” *Computer* 49 (2016): 3. 10.1109/MC.2016.373.

Quality of Service Optimization in Mobile Edge Computing Networks via Deep Reinforcement Learning

Li-Tse Hsieh[1], Hang Liu[1], Yang Guo[2], Robert Gazda[3]

[1] The Catholic University of America, Washington, DC 20064, USA
[2] National Institute of Standards and Technology, Gaithersburg, MD 20878, USA
[3] InterDigital Communications, Inc., Conshohocken, PA 19428, USA

Abstract. Mobile edge computing (MEC) is an emerging paradigm that integrates computing resources in wireless access networks to process computational tasks in close proximity to mobile users with low latency. In this paper, we propose an online double deep Q networks (DDQN) based learning scheme for task assignment in dynamic MEC networks, which enables multiple distributed edge nodes and a cloud data center to jointly process user tasks to achieve optimal long-term quality of service (QoS). The proposed scheme captures a wide range of dynamic network parameters including non-stationary node computing capabilities, network delay statistics, and task arrivals. It learns the optimal task assignment policy with no assumption on the knowledge of the underlying dynamics. In addition, the proposed algorithm accounts for both performance and complexity, and addresses the state and action space explosion problem in conventional Q learning. The evaluation results show that the proposed DDQN-based task assignment scheme significantly improves the QoS performance, compared to the existing schemes that do not consider the effects of network dynamics on the expected long-term rewards, while scaling reasonably well as the network size increases.

Keywords: Mobile Edge Computing (MEC), Task Assignment, Double Deep Q Networks (DDQN)

1 Introduction

The rapid development of Internet of Things (IoT) has generated a huge volume of data at the edge of the network. This requires a large amount of computing resources for big data analysis and processing, the capability of real-time remote control over both real and virtual objects, as well as physical haptic experiences. Cloud computing has been proposed as a promising solution to meet the fast-growing demand for IoT applications and services. However, centralized cloud data centers are often far from the IoT devices and users. How to provide high quality of service (QoS) to the interactive IoT applications, especially at the edge of the network, is still an open problem. This motivates a new paradigm referred to as mobile edge commuting (MEC), also called multi-access edge computing or fog computing, which extends cloud computing to the network edge

This work is partially supported by the National Science Foundation under Grants CNS-1910348 and CNS-1822087, and InterDigital Communications, Inc.

Hsieh, Li-Tse; Liu, Hang; Guo, Yang; Gazda, Robert. "Quality of Service Optimization in Mobile Edge Computing Networks via Deep Reinforcement Learning." Paper presented at The 15th International Conference on Wireless Algorithms, Systems, and Applications (WASA 2020), Qingdao, CN. September 13, 2020 - September 15, 2020.

[1, 2]. Edge nodes or edge devices provide computing services and carry out computationally intensive application and data processing tasks at the edge of the network between end users and cloud data centers. They can be computing servers or micro data centers deployed with routers, gateways, and access points in wireless access networks, and can also correspond to portable devices such as mobile phones, drones, robots, and vehicles with excessive computing resources that can be utilized to offer services to others. MEC can reduce transmission latency and alleviate network congestion. It also allows network operators to provide value-added real-time services and enhance QoS to end users.

A resource demand estimation and provisioning scheme for an edge micro data center is presented in [3] to maximize resource utilization. In [4], the authors proposed a hierarchical game framework to model the interactions where the edge nodes help the cloud data center operators process delay-sensitive tasks from mobile users and to determine the edge node resource allocation, service price, and pairing of edge nodes and data center operators with Stackelberg game and matching theory. These works focus on the interaction between edge nodes and cloud data centers to better serve the users, but they either abstract the MEC layer as a single edge server or assume that the edge nodes are independent of each other without consideration of their cooperation in processing tasks. The authors in [5] proposed an offloading scheme that allows a MEC edge node to forward its tasks to its neighboring edge nodes for execution to balance the workload fluctuations on different nodes and reduce the service delay. However, the paper made many idealized assumptions in assigning the tasks to the edge nodes, such as a fixed task arrival rate at each edge node as well as pre-known queuing delay of each node and transmission delay between the nodes. Their task assignment algorithm utilizes the classical model-based techniques that relies on these idealized assumptions to minimize service delay for one-shot optimization under a given deterministic MEC network state. Such an approach fails to capture the broad range of network parameters and ignores the impacts of dynamic network situations and heterogeneous nodes to the network performance.

On the other hand, reinforcement learning techniques can capture a wide range of control parameters and learn the optimal action, i.e. the policy for task assignments, with no or minimal assumptions on the underlying network dynamics. The conventional Q-learning algorithm is based on a tabular setting with high memory usage and computation requirements and is known to overestimate action values under certain conditions [6]. Recently, double deep Q networks (double DQN or DDQN) were introduced to address the problems of conventional Q-learning, which combines double Q-learning with two deep neural networks [7]. DDQN can provide large-scale function approximation with a low error and reduces the overestimations.

In this paper, we propose an online DDQN-based algorithm for task assignment in dynamic MEC networks, which accounts for both performance and complexity. The proposed algorithm takes into consideration the cooperation among the edge nodes as well as the cooperation between the edge nodes and a cloud data center. It performs sequential task assignment decisions in a series of control epochs to enable the nodes to help each other process user tasks and optimize a long-term expected QoS reward in terms of the service delay and task drop rate. The algorithm is designed to operate under stochastic and time-varying task arrivals, node processing capabilities, and network communication delays without a prior knowledge of these underlying dynamics. A decomposition technique is also introduced to reduce computational complexity in DDQN learning.

The remainder of the paper is organized as follows: Section 2 describes the problem formulation. In Section 3, we derive the online DDQN-learning based cooperative MEC task assignment algorithm in detail. In Section 4, we provide the numerical experimental results. Finally, the conclusions are given in Section 5.

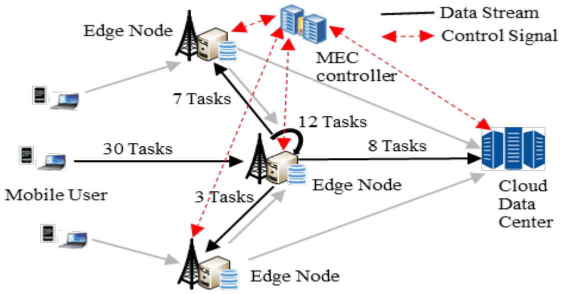


Fig. 1. An example MEC system model.

2 Problem Formulation

Fig. 1 illustrates an example MEC system model for consideration in this paper. A set of N edge nodes, $\mathcal{N} = \{1, \dots, N\}$, with computing, storage, and communication resources are co-located or integrated with cellular base stations (BSs) or WiFi access points (APs) in a wireless access network. IoT devices or mobile users connect to nearby edge nodes through their cellular or WiFi radios and send their computation-intensive tasks to the edge nodes to be processed. When an edge node receives tasks from its associated users, it either processes them locally, or forwards part or all of its unprocessed tasks to other edge nodes or to a remote cloud data center for processing if the node does not have sufficient resources to complete all the tasks. The remote cloud data center, n_c is modeled as a special node that is equipped with powerful computing capability but incurs a high network delay due to the distant location.

We assume that the system operates over discrete scheduling slots of equal time duration. At the beginning of a time slot t, a controller in the MEC network collects the network conditions and determines a task assignment matrix, $\boldsymbol{\Phi}^t = [\phi_{n,j}^t : n, j \in \mathcal{N} \cup n_c\}$. It informs the edge nodes to offload or receive computing tasks to/from the other nodes depending on the task assignment, where $\boldsymbol{\phi}_n^t = [\phi_{n,j}^t, \phi_{j,n}^t : j \in \mathcal{N} \cup n_c\}$ represents the task assignment vector regarding edge node n. $\phi_{n,j}^t$ specifies the number of tasks that edge node n will send to node j for processing in the time slot t, and $\phi_{n,n}^t$ is the number of tasks that are processed locally by edge node n. We assume that the data center n_c will process all the received tasks by itself without offloading them to the edge nodes, i.e. $\phi_{n_c,j}^t = 0, j \in \mathcal{N}$.

We first formulate the problem of stochastic task assignment optimization and then explore the methods to solve the optimization problem. Each edge node maintains a queue buffering the tasks received from its users, and q_n^t represents the queue length of node *n* at the beginning of time slot *t.* The queue size is bounded as $q_n^{(max)}$. It is assumed that the number of computational tasks arrived at edge node *n* in time slot *t*, A_n^t, is random and its distribution is unknown in advance. We denote $\boldsymbol{A^t}$= $\{A_n^t : n \in \mathcal{N}\}$. The task processing capability of node *n* in time slot *t*, denoted as s_n^t, which is the maximal number of tasks that node *n* can serve in the slot *t*, is also time-varying and unknown in advance due to the variable task complexity and adaptation of CPU cycles based on the power status and heat. The queue evolution of node *n* can then be written as $q_n^{t+1} = \max\{0, \min[q_n^t + \mathrm{A}_n^t + \sum_{i \in e_n} \phi_{i,n}^t - \phi_{n,i}^t - s_{n,}^t, q_n^{(max)}]\}$, where $\sum_{i \in e_n} \phi_{n,i}^t$ with $e_n = \{\mathcal{N} \cup n_c\} \backslash \{n\}$ represents the number of tasks that edge node *n* offloads to other nodes, and $\sum_{i \in e_n} \phi_{i,n}^t$ is the number of tasks that edge node *n* receives from other nodes in slot *t*.

When an edge node *n*, $n \in \mathcal{N}$ offloads a task to another node j, $j \in \mathcal{N} \cup n_c$ for execution at time slot *t*, it incurs an network delay cost, denoted as $c_{n,j}^t$. Let $\boldsymbol{c}_n^t = (c_{n,j}^t, c_{j,n}^t : j \in \mathcal{N} \cup n_c)$ represent the network delay vector for offloading the tasks from node *n* to any other node j, or vice versa, and $c_{n,n}^t = 0$. The network delay between two nodes is also time-varying and unknown in advance due to dynamic network conditions, traffic load, and many other uncertain factors. For a node *n*, $n \in \mathcal{N} \cup n_c$, at the beginning of time slot *t*, we characterize its state by its queue size q_n^t, its task processing capability $s_{n,}^t$, and the delay cost to offload a task to other nodes $\boldsymbol{c}_n^t$, thus $\boldsymbol{\chi}_n^t = (q_n^t, s_{n,}^t \boldsymbol{c}_n^t)$. The global state of the MEC network at the beginning of scheduling slot *t* can be expressed as $\boldsymbol{\chi^t}$= $(\boldsymbol{\chi}_n^t: n \in \mathcal{N} \cup n_c) = (\boldsymbol{q^t}, \boldsymbol{s^t}, \boldsymbol{c^t}) \in X$, where $\boldsymbol{q^t}$= $\{q_n^t : n \in \mathcal{N} \cup n_c\}$, $\boldsymbol{s^t} = \{s_n^t : n \in \mathcal{N} \cup n_c\}$, $\boldsymbol{c^t} = \{\boldsymbol{c}_n^t : n \in \mathcal{N} \cup n_c\}$, and X represents the whole MEC system state space.

We consider real-time interactive IoT applications and employ the task service delay and task drop rate to measure the system QoS. The task service delay, d_n^t, is defined as the duration from the time a task arrives at an edge node to the time it is served, and the task drop rate, o_n^t, is defined as the number of dropped tasks per unit of time. Given the MEC network state, $\boldsymbol{\chi^t}$= $(\boldsymbol{q^t}, \boldsymbol{s^t}, \boldsymbol{c^t})$ at the beginning of a time slot *t*, a task assignment $\boldsymbol{\Phi}^t = \boldsymbol{\Phi}(\boldsymbol{\chi^t}) = [\phi_{n,j}(\boldsymbol{\chi^t}): n, j \in \mathcal{N} \cup n_c\}$ is performed, which results in an instantaneous QoS reward. We define the instantaneous QoS reward at time slot *t* as

$$U\big(\boldsymbol{\chi^t}, \boldsymbol{\Phi}(\boldsymbol{\chi^t})\big) = \textstyle\sum_{n \in \mathcal{N}} [w_d U_n^{(d)}(\boldsymbol{\chi^t}, \boldsymbol{\Phi}(\boldsymbol{\chi^t})) + w_o U_n^{(o)}(\boldsymbol{\chi^t}, \boldsymbol{\Phi}(\boldsymbol{\chi^t}))], \tag{1}$$

where $U_n^{(d)}(.)$ and $U_n^{(o)}(.)$ measure the satisfaction of the service delay and task drop rate, respectively. w_d and w_o are the weight factors indicating the importance of delay and task drop in the reward function of the MEC system, respectively.

As mentioned before, the task arrivals and network states are non-deterministic and vary over time. We therefore want to cast the task assignment as a dynamic stochastic optimization problem, which maximizes the expected long-term QoS reward of an MEC network while ensuring the service delay and task drop rate are within their respective acceptable thresholds. More specifically, we define $V(\boldsymbol{\chi}, \boldsymbol{\Phi}) = \mathrm{E}[(1 - \gamma) \sum_{t=1}^{\infty} \gamma^{t-1} U(\boldsymbol{\chi}^{\boldsymbol{t}}, \boldsymbol{\Phi}(\boldsymbol{\chi}^{\boldsymbol{t}})) |\, \boldsymbol{\chi}^{\mathbf{1}}]$ as the discounted expected value of the long-term QoS reward of an MEC network, where $\gamma \in [0, 1)$ is a discount factor that discounts the QoS rewards received in the future, and $(\gamma)^{t-1}$ denotes the discount to the (t −1)-th power. $\boldsymbol{\chi}^{\mathbf{1}}$ is the initial network state. $V(\boldsymbol{\chi}, \boldsymbol{\Phi})$ is also termed as the state value function of the MEC network in state $\boldsymbol{\chi}$ under task assignment policy $\boldsymbol{\Phi}$. Therefore, the objective is to design an optimal task assignment control policy $\boldsymbol{\Phi}^*$ that maximizes the expected discounted long-term QoS reward, that is,

$$\boldsymbol{\Phi}^* = \arg \max_{\boldsymbol{\Phi}} \big(V(\boldsymbol{\chi}, \boldsymbol{\Phi})\big) \tag{2}$$

$$\text{subject to } d_n^t \le d^{(\max)}, o_n^t \le o^{(\max)}, \forall: n \in \mathcal{N} \cup n_c$$

where $d^{(\max)}$ and $o^{(\max)}$ are the maximal tolerance thresholds for the service delay and the task drop rate, respectively. $V^*(\boldsymbol{\chi}) = V(\boldsymbol{\chi}, \boldsymbol{\Phi}^*)$ is the optimal state value function. We assume that the probability of a network state in the subsequent slot depends only on the state attained in the present slot and the control policy, i.e. the MEC network state $\boldsymbol{\chi}^{\boldsymbol{t}}$ follows a controlled Markov process across the time slots. The task assignment problem can then be formulated as a Markov decision process (MDP) with the discounted reward criterion, and the optimal task assignment control policy can be obtained by solving the following Bellman's optimality equation [8],

$$V^*(\boldsymbol{\chi}) = \max_{\boldsymbol{\Phi}} \{(1 - \gamma)\, U(\boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})) + \gamma \textstyle\sum_{\boldsymbol{\chi}'} \Pr\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\}\, V^*(\boldsymbol{\chi}')\}, \tag{3}$$

where $\boldsymbol{\chi}' = \{\boldsymbol{q}', \boldsymbol{s}', \boldsymbol{c}'\}$ is the subsequent MEC network state, and $\Pr\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\Phi}(\boldsymbol{\chi})\}$ represents the state transition probability to the next state $\boldsymbol{\chi}'$ if the task assignment $\boldsymbol{\Phi}(\boldsymbol{\chi})$ is performed in state $\boldsymbol{\chi}$. $\boldsymbol{q}' = \{q'_n : n \in \mathcal{N} \cup n_c\}$, $\boldsymbol{s}' = \{s'_n : n \in \mathcal{N} \cup n_c\}$, and $\boldsymbol{c}' = \{\boldsymbol{c}'_n : n \in \mathcal{N} \cup n_c\}$ are the queue, task processing capability, and network delay states in the subsequent time slot.

The traditional solutions to (3) are based on value iteration, policy iteration, and dynamic programming [9, 10], but these methods require a full knowledge of the network state transition probabilities and task arrival statistics that are unknown beforehand in our dynamic network case. Thus, we seek the online reinforcement learning approach which does not have such a requirement. In previous research, we introduced an algorithm based on conventional Q-learning [6], which defines an evaluation function, called Q function, $Q(\boldsymbol{\chi}, \boldsymbol{\Phi}) = (1 - \gamma)\, U(\boldsymbol{\chi}, \boldsymbol{\Phi}) + \gamma \sum_{\boldsymbol{\chi}'} \Pr\{\boldsymbol{\chi}' | \boldsymbol{\chi}, \boldsymbol{\Phi}\}\, Q(\boldsymbol{\chi}', \boldsymbol{\Phi})$ and learns an optimal state-action value table in a recursive way to decide the optimal task assignment control policy for each time slot. However, for the cooperative MEC network, the task assignment decision-making for a node depends on not only its own resource availability and queue state, but also is affected by the resource availabilities and queue states of other nodes. The system state space and control action space grows rapidly as the number of involved nodes increases. The conventional tabular-based Q-learning process will search and update a large state-action value table, which incurs high memory usage and computation complexity.

3 Optimal Task Assignment Scheme Based on DDQN

In this section, we focus on developing an efficient algorithm to approach the optimal task assignment policy based on recent advances in deep reinforcement learning, which combines Q-learning and deep neural networks to address the state and action space explosion issue of the conventional Q learning with no requirement for a prior statistical knowledge of network state transitions and user task arrivals. Specifically, we design a DDQN-based algorithm to approximate the optimal state value function. In addition, it can be observed that the QoS reward function is of an additive structure, which motivates us to linearly decompose the state value function, and incorporate the decomposition technique into the deep reinforcement learning algorithm to lower its complexity.

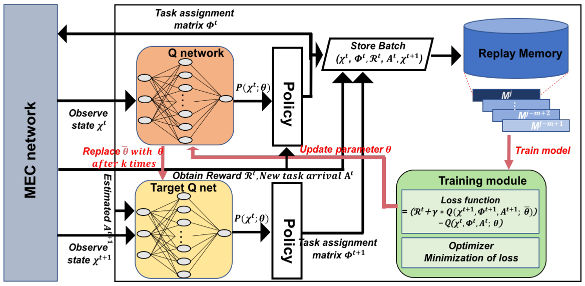

Task assignment matrix Φ^t
MEC network
Q network
$P(\chi^t;\theta)$
Policy
Store Batch $(\chi^t, \Phi^t, \mathcal{R}^t, A^t, \chi^{t+1})$
Replay Memory
M^j
M^{j-m+2}
M^{j-m+1}
Observe state χ^t
Replace $\bar{\theta}$ with θ after k times
Update parameter θ
Train model
Obtain Reward $\mathcal{R}^t$, New task arrival A^t
Estimated A^{t+1}
Target Q net
$P(\chi^t;\theta)$
Policy
Training module
Loss function $= (\mathcal{R}^t + \gamma * Q(\chi^{t+1}, \Phi^{t+1}, A^{t+1}; \bar{\theta})) - Q(\chi^t, \Phi^t, A^t; \theta)$
Optimizer Minimization of loss
Observe state χ^{t+1}
Task assignment matrix Φ^{t+1}

Fig. 2. DDQN-based cooperative MEC task assignment.

Fig. 2 illustrates the DDQN-based reinforcement learning scheme for the collaborative MEC task assignment. DDQN replaces the tabular setting of conventional Q-learning with two neural networks, Q evaluation network and Q target network, to learn and approach the optimal state value function and decide the optimal action [7]. The Q evaluation network (Q-eval) is used to select the task assignment matrix $\boldsymbol{\Phi}^{\boldsymbol{t}}(\boldsymbol{\chi}^{\boldsymbol{t}};\theta)$ based on the collected network states $\boldsymbol{\chi}^{\boldsymbol{t}}$ at the time slot *t*, and the Q target network (Q-tar) is used to select the task assignment matrix $\boldsymbol{\Phi}^{\boldsymbol{t+1}}(\boldsymbol{\chi}^{\boldsymbol{t+1}};\bar{\theta})$ at the following time slot *t*+1. The parameters θ and $\bar{\theta}$ can be learned and updated iteratively. The standard DDQN algorithm outputs the state-action values and select the action with the maximum Q value. Unfortunately, the traditional DDQN approach in [7] cannot be directly applied to solve our problem because we do not know the number of the new task arrivals in a time slot at the beginning of that time slot. To solve the problem, we modified the Q-eval and Q-tar networks in the standard DDQN to output a probability matrix, which indicates the probability to forward a task from one edge node to another edge node in the slot.

The modified DDQN is used to approximate the optimal state value function in (3) and select the best action. We redefine the state value function (3) as

$$V^t(\boldsymbol{\chi}^{\boldsymbol{t}}) = \max_{\boldsymbol{\Phi}}\{(1-\gamma^t)U(\boldsymbol{\chi}^{\boldsymbol{t}}, \boldsymbol{\Phi}^{\boldsymbol{t}}(\boldsymbol{\chi}^{\boldsymbol{t}}, \mathcal{P}(\boldsymbol{\chi}^{\boldsymbol{t}}; \theta^t))) +$$
$$\gamma^t[\Pr\{\boldsymbol{\chi}^{\boldsymbol{t+1}}|\boldsymbol{\chi}^{\boldsymbol{t}}, \boldsymbol{\Phi}^{\boldsymbol{t}}(\boldsymbol{\chi}^{\boldsymbol{t}}, \mathcal{P}(\boldsymbol{\chi}^{\boldsymbol{t}}; \theta^t))\}U(\boldsymbol{\chi}^{\boldsymbol{t+1}}, \boldsymbol{\Phi}^{\boldsymbol{t+1}}(\boldsymbol{\chi}^{\boldsymbol{t+1}}, \mathcal{P}'(\boldsymbol{\chi}^{\boldsymbol{t+1}}; \bar{\theta}^t)))]\}, \tag{4}$$

where $\mathcal{P}(\boldsymbol{\chi}^{t};\theta^{t})$ and $\mathcal{P}'\left(\boldsymbol{\chi}^{t+1};\bar{\theta}^{t}\right)$ are the probability matrices calculated by Q evaluation and Q target networks, respectively. In the standard DDQN algorithm, the state value will be updated in each time slot and used to determine the optimal action. To simplify the updates, in our implementation, the state value obtained from (4) is stored in a replay memory for training and updating θ and $\bar{\theta}$ in the learning process so that the Q-eval and Q-tar can select the optimal task assignment matrices directly and accurately. The loss function for updating the parameters θ of Q-eval can be defined as

$$\mathbb{L}(\theta) = E\left[\left((1-\gamma)U(\chi,\boldsymbol{\Phi}(\chi,\mathcal{P}(\chi;\theta))) + \gamma U(\chi',\boldsymbol{\Phi}'(\chi',\mathcal{P}'(\chi';\bar{\theta}))) - V(\chi)\right)^{2}\right], \tag{5}$$

and the parameters $\bar{\theta}$ will be updated by copying θ after a predefined number of steps.

At the beginning of each time slot *t*, the MEC controller determines the task assignment matrix $\boldsymbol{\Phi}^{t}(\boldsymbol{\chi}^{t})$ based on the collected network states and informs the edge nodes of the task assignment decision. The task assignment matrix $\boldsymbol{\Phi}^{t} = [\phi_{n,j}^{t}: n, j \in \mathcal{N} \cup n_{c}\}$ at the beginning of scheduling slot *t* is determined as,

$$\boldsymbol{\Phi}^{t} = \mathcal{P}^{t}(\chi^{t};\theta^{t}) \tag{6}$$

An edge node then offloads the tasks to other nodes or receives tasks from other nodes and processes these tasks based on the task assignment decision. The new task arrivals $\boldsymbol{A^{t}}$ will be counted at the end of the time slot t and the new network state is collected and updated to $\boldsymbol{\chi^{t+1}}$ by the controller. The MEC network receives a QoS reward $\boldsymbol{U}^{t} = U(\boldsymbol{\chi^{t}},\ \boldsymbol{\Phi^{t}}(\chi^{t},\mathcal{P}(\chi^{t};\theta^{t})))$ by performing the task processing. The Q-tar network is used to calculate $\boldsymbol{\Phi^{t+1}}$. As mentioned before, the DDQN includes a replay memory that is used to store a pool of the most recent M transition experiences, $\Omega^{t} = \{\boldsymbol{m}^{t-M+1},\dots,\boldsymbol{m}^{t}\}$, where each experience $\boldsymbol{m}^{t} = (\boldsymbol{\chi}^{t},\boldsymbol{\Phi}^{t},\boldsymbol{U}^{t},\boldsymbol{\chi}^{t+1},\boldsymbol{\Phi}^{t+1})$ is occurred at the transition of two consecutive slots *t* and *t* + 1 during the learning process. At each slot *t*, the *k* previous experiences are randomly sampled as a batch from the memory pool Ω^{t} to train the DDQN online. The learning process will calculate the approximated overall state value for each experience in the batch and update the parameters θ with an goal to minimize the loss function (5). Once the state value function is converged, we can obtain the optimal parameters θ^{*} for Q-eval. The optimal policy will be

$$\boldsymbol{\Phi}^{*} = \mathcal{P}^{*}(\chi;\theta^{*}) \tag{7}$$

The MEC network QoS reward in (1) is the summation of the service delay and task drop rate satisfactions of the edge nodes, and the task arrival statistics and task processing capabilities of the edge nodes are independent each other. We can then decompose (4) into per node QoS reward and separate the satisfactions regarding the service delay and the task drops [11]. We first rewrite (6) as

$$\boldsymbol{\Phi}^{t} = \boldsymbol{\Phi}^{t}(\chi^{t}) = \{\phi_{n}^{t}(\chi_{n}^{t}): n \in \mathcal{N}\}. \tag{8}$$

n agents $n \in \mathcal{N}$ can be employed and each agent learns the respective optimal state value function through a per node sub-DDQN. An optimal joint task assignment control decision is thus made to maximize the aggregated state value function from all the agents. The task assignment related to node *n* can be expressed as

$$\phi_n^t(\chi_n^t) = \mathcal{P}_n^t(\chi_{\boldsymbol{n}}^{\boldsymbol{t}}; \theta_n^t), \tag{9}$$

where $\mathcal{P}_n(.)$ is the task assignment probability obtained through DDQN n. The state value function in (4) can be decomposed and expressed as in (10) and (11)

$$V^t(\boldsymbol{\chi}^t) = \ \sum_{n\in\mathcal{N}} V_n^t(q_n^t, s_n^t, c_n^t), \tag{10}$$

$$V_n^t(\chi_n^t) = \ (1-\gamma^t)U(\chi_n^t,\ \Phi^t(\chi_n^t,\ \mathcal{P}_n(\chi_n^t;\theta_n^t))) + \\ \gamma^t\left[\Pr\{\chi_n^{t+1}|\chi_n^t,\ \Phi^t(\chi_n^t,\ \mathcal{P}_n(\chi_n^t;\theta_n^t))\}U\left(\chi_n^{t+1},\Phi^{t+1}(\chi_n^{t+1},\mathcal{P}_n'\left(\chi_n^{t+1};\overline{\theta}_n^t\right))\right)\right] \tag{11}$$

With the linear decomposition, the problem to solve a complex Bellman's optimality equation (4) is broken into simpler MDPs and the computation complexity is lowered. In order to derive a task assignment policy based on the global MEC network state, $\boldsymbol{\chi} = (\chi_n : n \in \mathcal{N} \cup n_c)$ with $\chi_n = (q_n, s_n, c_n)$ and $c_n = \left(c_{n,j}, c_{j,n}\ :\ j \in \mathcal{N} \cup n_c\right)$, at least $\prod_{n\in\mathcal{N}\cup n_c}\prod_{j\in\mathcal{N}\cup n_c}(\,|q_n|\,|s_n||c_{n,j}||c_{j,n}|)$ states should be trained. Using linear decomposition, only $(N+1)\ |q_n||s_n|\ \prod_{j\in\mathcal{N}\cup n_c}(|c_{n,j}||c_{j,n}|)$ states need to be trained, resulting in much simplified task assignment decision makings and significantly reducing training time. The online DDQN-based algorithm to estimate the optimal state value function and determine the optimal task assignment policy is summarized in Algorithm 1.

Algorithm 1. Online DDQN-based Cooperative MEC Task Assignment

```
1.  Initialize the Q-eval and the Q-tar with two sets of $\theta^t$ and $\bar{\theta}^t$ random parameters for $t = 1$; the replay memory $M^t$ with a finite size of $M$ for experience replay.
2.  At the beginning of scheduling slot $t$, the MEC controller observes the network state, $\boldsymbol{\chi}^t = \{\boldsymbol{\chi}_n^t : n \in \mathcal{N}\}$ with $\boldsymbol{\chi}_n^t = \ (q_n^t, s_n^t, \boldsymbol{c}_n^t)$, and the Q-eval with parameters $\theta^t$ determines the task assignment matrix, $\boldsymbol{\Phi}^t = \ [\boldsymbol{\phi}_n^t : n \in \mathcal{N}]$ according to (8) and (9).
3.  After offloading and processing the tasks according to the above task assignment decision, the edge nodes will receive new tasks $\boldsymbol{A}^t = \ \{A_n^t\ : n \in \mathcal{N}\}$ at the end of slot $t$.
4.  The controller determines the QoS reward $U^t$after new task arrivals and calculates the state value $V^t$ according to (10) and (11)
5.  The network state transits to $\boldsymbol{\chi}^{t+1} = \{\boldsymbol{\chi}_n^{t+1} : n \in \mathcal{N}\}$ where $\boldsymbol{\chi}_n^{t+1} = \ (q_n^t + A_n^t, s_n^{t+1},\ \boldsymbol{c}_n^{t+1})$, which is taken as input to the target DQN with parameter $\bar{\theta}^t$ to select task assignment matrix $\boldsymbol{\Phi}^{t+1} = \{\boldsymbol{\phi}_n^{t+1}, n \in \mathcal{N}\}$ at the following scheduling slot $t$+1.
6.  The replay memory $M^t$ is updated with most recent transition $\boldsymbol{m}^t(\boldsymbol{\chi}^t,\ \boldsymbol{\phi}^t,\ \boldsymbol{U}^t,\ \boldsymbol{\phi}^{t+1}, \boldsymbol{\chi}^{t+1})$.
7.  Once the memory replay collect $\overline{M}$ transitions, the controller updates the Q-eval parameter $\theta^t$ with a randomly sampled batch of transitions to minimize (5)
8.  The target DQN parameters $\bar{\theta}^t$ are reset every $k$ time slots, and otherwise $\bar{\theta}^t = \bar{\theta}^{t-1}$
9.  The scheduling slot index is updated by $t \leftarrow t + 1$.
10. Repeat from step 2 to 9.
```

4 Numerical Experiments

In this section, we evaluate the cooperative MEC task assignment performance achieved by our derived online DDQN-based algorithm. Throughout the simulation experiments, we assume that the processing capability s_n^t, $\forall n \in \mathcal{N}$ of different edge nodes are independent of each other and evolve according to a Markov chain model, each modeled with three states characterizing the high, medium, and low with {4, 2, 1} tasks per slot. We simulated multiple MEC network scenarios with different system parameters. Due to the page limit, we present the results for several typical settings. The slot duration is set to be 30 msec. The network delay between two edge nodes, c_{nj}^t, $\forall n, j \in \mathcal{N}$, is also modeled as a Markov chain with three states, {1, 0.5, 0.2} slots. Edge nodes communicate with a cloud data center through the Internet. The network delay between the edge node and the cloud data center $c_{nn_c}^t$, $\forall n \in \mathcal{N}$ is assumed to be 10 slots. $U_n^{(d)}$ and $U_n^{(o)}$ in the QoS reward function are chosen to be the exponential functions [12] with $U_n^{(d)} = \exp\left(-d_n^t/d^{(\max)}\right)$ and $U_n^{(o)} = \exp\left(-o_n^t/o^{(\max)}\right)$.

The neural networks used for Q-tar and Q-eval have a single hidden layer with 15 neurons. We use ELU (Exponential Linear Unit) as the activation function for the hidden layer and Softmax for the output layer to output the probability matrices for the action selection. The optimizer is based on RMSProp [7]. The number of iterations for updating parameters of Q-tar is set to be 30, and the memory replay size and the batch size are also set to be 30. The training process will be triggered when the system collects enough samples and it will pull out all samples to train. There are other sampling optimization techniques, e.g. prioritized experience replay, which will be included in our future work.

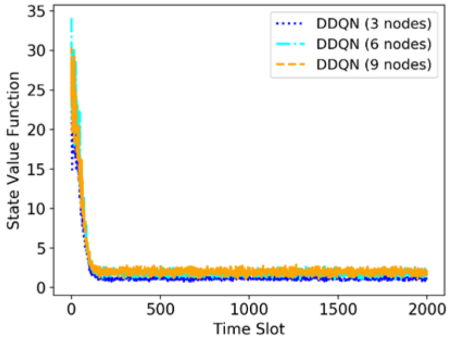


Fig. 3. Convergence of the proposed DDQN-based learning process.

We first investigate the convergence performance of the proposed online DDQN-based cooperative MEC task assignment algorithm under dynamic stochastic MEC network environments with different number of MEC edge nodes. As shown in Fig. 3, we can observe that the proposed algorithm spends a short time period to learn and then converges to the global optimal solution at a reasonable time period which is less than

150 slots. In addition, the network size does not have noticeable effects on the convergence time of the algorithm.

Next, we evaluate the QoS performance of the proposed online DDQN-based cooperative task assignment scheme. For the purpose of comparison, we simulate four baselines as well, namely,

1) No Cooperation: An edge node processes all the tasks it receives from its associated users by itself. There is no task offloading.
2) Cloud Execution: An edge node offloads all its received tasks to the cloud data center for execution.
3) One-shot Optimization: Like the scheme in [5], at each scheduling slot, the task assignment is performed with the aim of minimizing the immediate task service delay. Note that the power efficiency constraint is not considered here because we assume the edge nodes have sufficient power supply.
4) Q-Learning: Task assignment optimization based on conventional Q-learning.

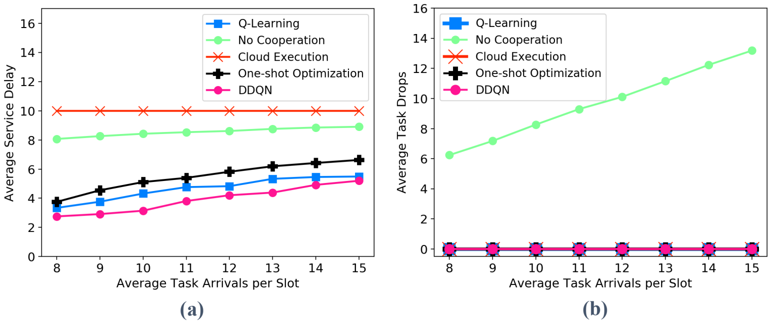


Fig. 4. (a) the average task service delay and (b) the average number of dropped tasks per slot versus the average task arrivals per slot for different algorithms.

Figures 4 (a) and (b) show the average task service delay and the average number of dropped tasks per slot, respectively, for the proposed scheme and baselines, with three edge nodes and one cloud data center as the task arrivals per slot at the edge nodes follow independent Poisson arrival process. The delay is measured in the unit of the time slot duration. We can observe that the DDQN-based and conventional Q-learning based task assignment schemes perform better than the other baselines such as No Cooperation, Cloud Execution, and One-shot Optimization schemes. This is because they not only consider the current task processing performance but also take into account the QoS performance in the future when determining the optimal task assignment matrix under time-varying stochastic task arrivals and network states. Their task drops are zero because the algorithms tend to minimize the task drops, and the edge nodes will forward the tasks to the cloud data center when their buffers are full. For the No Cooperation scheme, an edge node does not send the unprocessed tasks to the cloud and other edge nodes, so that there are tasks drops when the node's buffer becomes overflow. For the Cloud Execution scheme, a large network delay is always incurred to ship the tasks to the cloud data center for processing over the Internet. The One-Shot Optimization scheme performs relatively well. However, it makes task assignment decisions to minimize the immediate task service delay in a slot and may cause shipping many tasks to the cloud data center for processing under fluctuating task arrivals and non-stationary node process capabilities, with such tasks incurring a large network delay.

Fig. 5 shows the memory usage of DDQN- and Q-learning task assignment schemes. The traditional tabular Q-learning consumes much higher system resources than the DDQN scheme and cannot scale well due to the explosion in state and action spaces, making the solution unviable. On the other hand, the memory usage by the DDQN-based task assignment scheme scales well as the number of edge nodes in the network increases.

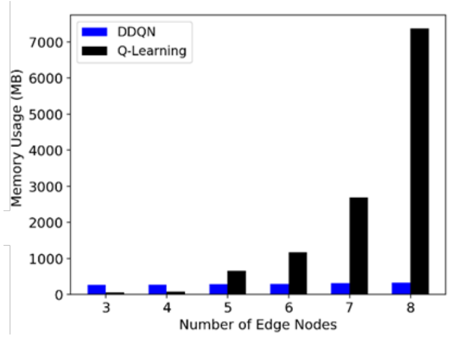


Fig. 5. The memory usage of DDQN and Q-learning task assignment schemes.

5 Conclusions

In this paper, we have investigated the task assignment problem for cooperative MEC networks, which enables horizontal cooperation between geographically distributed heterogeneous edge nodes and vertical cooperation between MEC edge nodes and remote cloud data centers to jointly process user computational tasks. We have formulated the optimal task assignment problem as a dynamic Markov decision process (MDP), and then proposed an online double deep Q-network based algorithm to obtain the optimal task assignment matrix. A function decomposition technique is also proposed to simplify the problem in DDQN learning. The proposed online DDQN algorithm does not require for a statistical knowledge of task arrivals and network state transitions. The evaluation results validate the convergence of the proposed algorithm and demonstrate that it outperforms the traditional schemes that optimize the immediate task service delay with no consideration of the impacts of network dynamics to the expected long-term QoS rewards. In addition, the proposed DDQN scheme can scale reasonably well, and requires much less memory than the conventional Q-learning based algorithm.

Acknowledgements

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

1. M. Patel, B. Naughton, C. Chan, N. Sprecher, S. Abeta, A. Neal, et al., “Mobile-edge computing introductory technical white paper," White Paper, Mobile-edge Computing (MEC) industry initiative, 2014.
2. H. Liu, F. Eldarrat, H. Alqahtani, A. Reznik, X. de Foy, Y. Zhang, "Mobile Edge Cloud System: Architectures, Challenges, and Approaches," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2495-2508, Sept. 2018.
3. M. Aazam and E.-N. Huh, “Dynamic resource provisioning through fog micro datacenter,” in Proc. of IEEE PerCom Workshops, pp. 105–110, St. Louis, MO, Mar. 2015,
4. H. Zhang, Y. Xiao, S. Bu, D. Niyato, F. R. Yu, and Z. Han, “Computing resource allocation in three-tier IoT fog networks: A joint optimization approach combining stackelberg game and matching,” IEEE Internet Things J., vol. 4, no. 5, pp. 1204–1215, 2017.
5. Y. Xiao and M. Krunz, “QoE and power efficiency tradeoff for fog computing networks with fog node cooperation,” in Proc. of IEEE INFOCOM’17, Atlanta, GA, May 2017.
6. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 1998.
7. H. van Hasselt, A. Guez, and D. Silver, “Deep reinforcement learning with double Q-Learning,” *In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (*AAAI'16*), Pages 2094–2100, February 2016.
8. D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, Belmont, MA, 1995.
9. M. L. Puterman and M. C. Shin, “Modified policy iteration algorithms for discounted Markov decision problems,” *Management Science*, vol. 24, no. 11, pp. 1127–1137, 1978.
10. R. Howard, *Dynamic Programming and Markov Processes*. MIT Press, 1960.
11. J. N. Tsitsiklis and B. van Roy, “Feature-based methods for large scale dynamic programming,” *Mach. Learn.*, vol. 22, no. 1-3, pp. 59 - 94, Jan. 1996.
12. X. Chen, Z. Zhao, C. Wu, M. Bennis, H. Liu, Y. Ji, and H. Zhang, “Multi-Tenant Cross-Slice Resource Orchestration: A Deep Reinforcement Learning Approach,” *IEEE Journal on Selected Areas in Communications* (*JSAC*), vol. 37, no. 10, pp. 2377 – 2392, Oct. 2019.

Balisage: ***The Markup Conference***

Systems security assurance as (micro) publishing

Declarative markup for systems description and assessment

Wendell Piez

Abstract

Markup technologies are very general purpose, as reflects their generality of conception. They become interesting as well as useful as they are applied to accomplish goals in the real world. Since principles of generic declarative markup were first applied to accomplishing publishing-related goals in information management, design and application, 25 or 40 years ago, they have repeatedly demonstrated both their generality – they really do work – and their demand for applicability. Get one thing wrong, or leave it out, and the effort sits on a shelf. Design and deploy it carefully and sensitively, and even an inexpensive initiative can pay dividends for years. These systems become sustainable in the context of the sustainable operations of which they are a part.

Decades of experience have shown us how to use declarative markup to sustain publishing operations. Now we have to deal with similar problems of information description, management, reuse across contexts, referencing, tracing, and authentication, only at even larger scales than before, both in size and complexity. This paper proposes some lessons and insights we can bring from our experience with publishing technologies, and suggests how they might be applicable in the growing domain of systems security assurance.

Table of Contents

Context of the conversation
- Practical successes of declarative markup
- What is publishing? challenges of documentation in systems security

What have we learned from technology in publishing?
- Your system is someone else's subsystem
- Declarative markup adds value
- Data acquisition is hard
- Activities are supported by incentive structures
- Quality is defined within context
- Evolution works by little revolutions

How do we know we need (something like) OSCAL?

The OSCAL Approach

Applying the lessons: conclusions and expectations

Context of the conversation

> ***Note***
>
> Disclaimer: Certain commercial equipment, instruments, or materials are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose. The opinions, recommendations, findings, and conclusions in this publication do not necessarily reflect the views or policies of NIST or the United States Government.

Practical successes of declarative markup

In the form of XML (Extensible Markup Language), for twenty years we have had systems exploiting the principles of declarative markup on a standards basis, with commodity tools. If you count XML's predecessors including both SGML[1] and applications of other technologies that are or can be “declarative” in their approach (such as LaTeX),[2] this history is much longer. This is no accident inasmuch as the roots of these technologies are in typesetting, among other requirements, which presents problems difficult enough to demand we factor out and “layer” solutions addressing challenges in functionality, configurability and maintainability. The layered solution to the problem of maintaining multiple publishing streams from complex aggregated sources, it turns out, is the layered solution to much else as well.

What has not happened? Whether considered as a standard processing stack based on the W3C (Worldwide

Web Consortium) XPath/XQuery Data Model (XDM xdm2017), or only as a data format, XML has not become the single and sufficient solution to all problems. Perhaps it was never, indeed, meant to be, at least not seriously – let us distinguish advocacy from marketing – yet in 2020, even as the broad domains of digital information stretch far beyond what visionaries of 20 years ago imagined, XML or even "layered", declarative approaches considered more generally, are by no means predominant. This is in part because other solutions have emerged to other problems, which have seemed more exigent or demanding than the long-term problems solved by descriptive markup or XML, and which do not entail its overhead. Those of us who work with it know that XML (or the more important principle it stands for) has not "failed"; yet at the same time, it has not altogether taken the field either. Sometimes its successes have been ambiguous. (War stories could be told.) It is probably closer to say that XML has been remarkably, even spectacularly successful in some ways – while in other ways, the future we imagined has not come to pass, or if it has, it appears in a form quite unlike what we expected. (People edit wikis. They use Markdown! War stories could be told.)

What is publishing? challenges of documentation in systems security

Into this ambiguous context we step with a new set of problems, unprecedented and yet (in many respects) familiar at the same time. Systems security assessment, assurance, authorization. Information exchange around systems security – this is not publishing exactly, in anything like the common sense (of making materials available to a public), yet it entails all the same problems of information gathering, organization, exposition, design and presentation for the consumption of readers and consumers, both sentient (people) and automated (machines).

In particular, systems security assessment or RMF-based security assurance activities[3] are not publishing in the sense that they entail creating productions for a general audience. Few of the documents produced by security professionals in their work are "published" in a normal sense. But the size of the audience, or even whether a document is released or disseminated to a public, are not the only defining features of what is "publishing". Perhaps we could refer to this (formal and informal) circulation of formatted office documents as "micropublishing" or targeted publishing. A PDF or Word document, that is, prepared at considerable trouble and expense by dedicated professionals, and submitted for review, whether it be by potential customers, regulatory authorities, or partner organizations, might never be "published", while it is nevertheless subject to all the same functional requirements in information creation, production, management, and tracking – and most especially, for revision cycles and quality control.

Compared to more normal sorts of publishing, this set of activities might work at a higher order of complexity, over faster – and also more extended! – time frames, with larger and more articulated information sets. Even how this is to be done adequately, much less at its best or in its ideal form, has hardly been defined, and its definition is itself dynamic, a moving target, as we learn more about systems and risk migitation and management. At the same time at a certain level we have no choice: this is work that will and must be done, at some level, the only question being how well. In this respect, "publishing" serves as shorthand for an entire range of activities entailing data collection, composition, analysis, review, formatting for presentation and finally the production of "artifacts" for consumption, be they "reports" or "proposals" or "reviews", "specifications" or "assessments" in whatever form - paper, PDF, Office or word-processor formats, web pages, or any other form. In these forms, these artifacts are disseminated to recipients that are able to make use of them for their own processes..

In a paper delivered to Balisage 2019, some of the special challenges of the (so-called) "system security" application domain – as seen from the point of view of a relative newcomer – were described at some length. Additionally, Joshua Lubell's companion paper to this one frames in much greater detail the questions of security assurance processes as a specifically *documentary* activity, and one in which (moreover) the questions of authority, knowledge, sources of knowledge, trust, accountability, traceability and transparency – all issues that have been implicit in every publishing system ever built (whether digital and networked, or by any analog media) – become especially salient. For purposes of this paper, this background is assumed as context.

Indeed, insofar as systems security may also entail both "marketing" and "customer relations" – even while it entails much else – we can recognize that even when it is not formalized to the extent that we see with the United States Federal Government's Risk Management Framework – as one approach to security among others – it is always based in the *appropriate transmission and communication of information between points*. In other words, it always comes down to a kind of publishing, albeit, as noted, a kind of targeted micropublishing. The details are always different. But a principled approach to the design of information technologies, which works to address one set of thorny problems in information exchange, should provide similar advantages when dealing with another.

What have we learned from technology in publishing?

Your system is someone else's subsystem

Self-similarity across scales - Information ecosystems or ecologies, as noted at Balisage 2018 (Piez 2018), are fractal in organization; one way we know this is by the way we find similar organizations and patterns of organization appearing at many levels of scale. In particular, with publishing workflows we can see a great range of scales and indeed nested scales. Every system is made of systems, with more or less articulated boundaries within the subsystems; and this is true of subsystems too. Essentially, all systems are hybrid systems, and the way to look at any one of them (whether at a "system" level or that of a component) is to consider its interfaces and externalities (including operational and technical dependencies) and especially the way these affect its capacity for maintenance, adaptation and scaling.

Consider for example the banal case of a scientific or scholarly journal. Its communications are carried forward for the sake of producing articles and making them available to a readership. Each article entails a complex choreography of data exchange, as the text of the article is submitted to the journal by its author, reviewed, revised (or rather: rejected in favor of revision), finally accepted and then "processed" (the means varies) for publication. Each of these steps entails one or more communications between parties to accomplish. These communications include the article itself but also the coordination and meta-commentary around it.

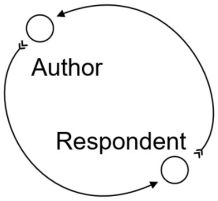
In principle this is measurable. A typical small journal might publish four or six issues per year; each of these issues might contain four or six articles, averaging 20 to 30 articles per year. Each of these articles requires a varying (small) number of peer reviews – which we can also count, yielding another number – say, between 50 and 100 peer reviews, per journal, per year. Circulating these peer reviews – and, more importantly, ensuring that documents are revised accordingly – constitutes much (though, not all) of the work of the journal editor perhaps with the help of staff. Since many or most peer reviews will then be returned to an author, for each peer review, there are at least three or four participants in the workflow (author, editor, peer reviewer, staff). To the extent there is attrition, as not all peer reviews and article revisions are carried through successfully, we could quantify this as well. In any case, even without numbers and figures (elapsed time per peer review, to edit it and return it to the author), it becomes clear how the work associated with such circulation is probably the largest limiting factor preventing a journal (on this model) from scaling up to a larger run. Essentially this means that due to the centrality of the editor in the worflow (if only as "peer review conductor") – there is a limit to the effective size of a journal. Making a journal bigger (in terms of production, not circulation) is much harder than spinning off a new journal.

One gating factor here is the relative difficulty of not just peer review, that is, but all the coordination around it. Author, editor, managing editors, peer reviewers: as long as they share no simple platform or standard for the handling of peer reviews, their peer reviewing system is essentially made up of the combination and intersection of all their personal systems, and maintaining communications this way is costly (even while regarded as normal and regular). Migrate peer reviewing to an online system, for example, that consolidates the effort of reviewing and tracking, and the numbers shift. Scaling is now easier – albeit other limiting factors such as availability or motivation, might remain.

Similarly, further along the lifecycle, once journals start being aggregated together, peer review is no longer a limiting factor. Once articles enter "post publication", they are more like black boxes, identifiable by metadata but not requiring intervention for particular cases (as peer review might). So publishers or aggregators of published information can pull together multiple issues of multiple journals, and the scaling bottleneck posed by peer reviews (or again more precisely, by the interventions they require), does not apply. Similarly – but differently – this particular bottleneck does not apply to other sorts of publishing such as trade or monograph publishing, where peer review and the revision process are done and managed differently.

These differences and distinctions between systems and the industries they serve, are driven and defined by their information processing requirements, and the difficulty and expense of those requirements – most of which are the expenses of time and expert attention. What we do not have, to take account of all this, is a science of workflows,[4] which is to say a set of principles and governing ideas for how workflows actually work: a branch of sociology and economics, but with a technical aspect insofar as workflows lend themselves to quasi-formal definition, once (for example) we start to specify the details of inputs and outputs.

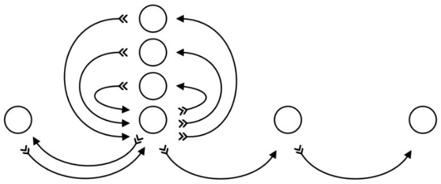
Figure 1: Elements of workflow



One principle of a science of worflow should be the concept of exchange. A party gives something to another party, who gives something in return. This response typically triggers another action, perhaps an iteration with a modified input. Such exchanges might be the "elements" or conceptual primitives of a systematic accounting.

When the parties are reciprocal and co-equal, and contents of the exchanges are equally contributed by both, we have a basic "Correspondence" pattern.

Figure 2: Peer review pattern



"Peer review" is one pattern in a (potential) pattern language to describe information workflows. Others might be "bundle"; "enhance" (e.g., provide metadata); "test" or "confirm"; "publish" (maybe both "push" and "pull") and so forth.

What a science of workflow would enable us to see would be the articulated joints of these related processes – where there are hand-offs in responsibility, and where there are requirements and capabilities for (what should be called) *transformations*, in that their outputs (results, what is "produced") are modifications, translations and enhancements of their inputs (their sources and raw materials). These can be and typically are very specific operations to very local sets of requirements. Nor is this a bug or a problem, in that these particularities are often the entire point of the exercise. (The reason we send an article to a peer reviewer, is that we would like to see what comes back.)

Without such a science, however, there are still things we can know from experience. One reason it is useful to work through a mundane example such as a peer review exercise in a hypothetic journal, is that it dramatizes the underlying reasons why, for example, rates of technological evolution are so uneven. It remains an open question, for example, in 2020, what format or formats are best used to share draft articles (or more precisely, the raw materials of what is to become an article) in a peer review process. Externalities (such as the ubiquity of certain tools or toolkits) push one way, while functional requirements – or even cultural considerations – in the system itself may push another. Thus there is no perfect solution – journals accept documents created with proprietary word processors because that is all their authors are prepared for. When an unusual consistuency is able to provide a journal with something better (perhaps a community of scientists, or academic scholars with text encoding skills, has tools they prefer) they will often take that. *This is all because the system of the journal is just a subsystem for each of its authors.*

In a fractal landscape, scales are relative, we should expect the same problems or versions of them to turn up in more than one place, albeit differently. While by definition and design, technologies of automation can support scaling, the fact that they also need to be *fitted* so closely – that two cases (say, two journals, or two research articles in a journal) are so similar without being alike – frustrates and even prohibits a cookie cutter approach.

What is true in one domain of information processing, might well be true in others. In publishing, especially technical publishing (however defined), things are always the same except where they are not. Precisely to deal with this variability, experience shows, only a well-defined, openly-specified, and non-proprietary technology can serve as the basis for (open-ended) “solutions”, which can be adapted to serve a heterogeneous and changing set of organizations, with their interlocking goals. Experience with academic, scientific and technical publishing indicates that such a technology will be declarative in form. For machine-readable data to be useful “for the duration” (that is the lifecycle of the information, not the system it currently sits in or the form it currently takes), it must systematically and consistently address and characterize the data itself as both artifact and “mechanism”. What that mechanism is or should be, is relative to the uses to which we put this information and our needs for handling and processing it.

Declarative markup adds value

As noted, it is possible for a scholar, researcher or “information professional” in 2020 to be entirely on a digital platform, where all the works apart from some odd print artifacts, consumed as well as produced, take digital form, and indeed in which XML, standards, declarative markup and open systems (perhaps broadly including wikis, git repositories, and assorted other forms of hybrid hypertext media on line) are central to the system, while word processors and other page-oriented tools are secondary.

But those are outliers, and almost everyone involved in “authoring” or “research” phases of publishing today still uses word processors, spreadsheets and document formats along with email – almost necessarily a one-off format with little potential for data reuse – for passing their information across systems. “Office documents” as we broadly call them, are the assumed basis for data interchange. Although they exist, journals or publishers that readily work with other kinds of data inputs, even nominally standard formats, are indeed quite rare, and the cases that do so are illustrative. Similarly, despite all the demonstrated advantages of “single source” publishing, publishers that produce anything but pages first (albeit in digitally encoded form, which is to say digital artifacts in PDF or Postscript®), are also quite rare; any web version or archive version is treated as a secondary production. The possible efficiencies to say nothing of the more outlandish potentials of an XML-first workflow, are simply not perceived to be worth it. And indeed they may not be, if the trouble, expense and disruption are certain while the gains are hypothetical.

On the other hand, to claim that markup technologies have had no impact would be to entirely misconstrue what has happened in 25-some years. While to some it may appear that XML's day has come and gone, it keeps coming back and proving its usefulness: indeed it might be said that the larger-scale activities now happening routinely in the publishing space enabling both access, and long-term stability, for collections of unprecedented size and complexity – all of these would be for practical purposes impossible without strategies of declarative (descriptive) markup and the principle of open lingua franca that can be established on its base. Even if markup technologies have not made their way explicitly into the work practices of writers, researchers and editors, it still cannot be said – most especially in the case of HTML (Hypertext Markup Language) and the web, but this is not the only case – that these technologies are not significant. Indeed, the reasons we invest energy in producing these representations (indexable, retrievable, massable, filterable, stylable – in ways their word processor documents are not), is because they prove to be so valuable.

To a great extent, this is because the principles themselves are sound. Of course there is nothing especially “XML” about cleanly layered separation, and an architecture that reflects and responds to the requirements of its users not to “paint” information to appear one way or another, but to expose and maintain the information *as information* – which interestingly we do by describing it. XML, with its associated technologies, is a means to this end.

Data acquisition is hard

Perhaps we all agree already that rich, clean declarative markup is by far the preferred form not only for archiving but for production. Even if we do (and I am not sure I do), the question remains of where that rich information set comes from: how does the XML get there in the first place? By XML, here, we mean of course not XML itself, but a particular kind of XML, as exemplified by documentary formats with descriptive tag sets. For reasons that will become clearer later, I also mean (paradoxically) an “XML” that is not XML at all (in that, as I will discuss, the JSON Javascript Object Notation] variants of OSCAL formats seeks to share the same advantages offered by the XML variants). (OSCAL, the Open Security Controls Assessment Language, is discussed further below.)

In the real world, the answer to the problem of how do we acquire the data to start is generally, with difficulty. When information sources are created and first transmitted as “office documents” (meaning any of a species of word processor, spreadsheet software, whether proprietary or open-standard), conversion into XML can be done (with care, by a skilled operator) “by hand”, or it can be semi-automated. Either way it will require additionally a skilled human operator for supervision, definition of data quality, assessment. Given these challenges, automation is expensive, becoming cost-effective only when rates of production are large enough – and rules are clear enough – to reward economies of scale.

In general, as well, while large-scale data conversions supported by externalized providers is a way of making XML, this approach is only affordable or sustainable for certain kinds of information. Systems security and assessment is actually characterized by a great heterogeneity of information formats, including data sets produced by machines as well as by people. This great heterogeneity, plus requirements for sensitive handling of private or confidential data, together make it difficult to outsource the task of data description to a third party or external provider. Data security questions aside (how do you shop for a conversion vendor to reformat your most sensitive and proprietary strategic information?), the combination of high data complexity and distinctiveness (at the levels of subdomain, markets and enterprise) may make for no “sweet spot” for outsourcing data conversions, across the domain of systems security. Partner organizations may need to be able to do this for themselves and each other, without always relying on external expertise.

Other methods of acquiring XML markup also present opportunities along with their own challenges. For example, we can bring XML tools such as structured editors earlier into the workflow; similarly, we can design systems with user interfaces (wizards, forms interfaces) that abstract the structure and its encoding away from the view. Both of these have the effect of providing support for the human operator (writer or editor), while permitting the information to be “XML native”; and the advantages of such system, where they can be used, can be considerable. *Where they can be used* here is the operative qualifier – since this is by no means everywhere: both development and deployment require a level of engagement with both goals and technical means. Generally speaking it is only the more agile and more technically-minded organizations that have been able to take advantage of such opportunities.

Yet there may also be a narrow and somewhat arduous path forward to structured data that goes *through* office documents, not around them. The key here may be templates, which already have the advantage of being the preferred method for much of the industry for capturing their information – largely because the promises of templates *are*, in many respects, the promises of structured data, while the deployment architecture (documents layered with so-called styles) is the same or similar, albeit in proprietary form. In some cases, it may be possible to design combinations of templates, rule sets, transformations and document validations (dynamic checks and feedback) that together can help with the job of data conversion, from a representation internal to the word processor, into an externalized form. For certain very regular and generalizable species or subspecies of documents, in certain organizations, this approach might serve as a useful accelerant to getting structured information into the mix. Word processor as structured editor.

While such a solution is possible, who is going to build the solution, for whose use? Will it have to be producers of the data themselves (which would demand an extraordinary combination of disparate skills), or can there be a market for such development and innovation? If the model of the third-party data conversion vendor does not serve for addressing the need here, what does?

Moreover, it can be expected that this question will remain acute until we have both adequate specifications for shared data description (to whatever level of “standard” possible), and working systems that respect and implement these standards. Until then, cumbersome data conversions will remain an impediment. Organizations who can insulate themselves at the boundaries, defining for their partners what the specifications of these formats will look like, will have an advantage.

This brings us to incentive structures.

Activities are supported by incentive structures

> “‘When I use a word,’ Humpty Dumpty said in rather a scornful tone, ‘it means just what I choose it to mean — neither more nor less.’
>
> ‘The question is,’ said Alice, ‘whether you can make words mean so many different things.’
>
> ‘The question is,’ said Humpty Dumpty, ‘which is to be master — that’s all.’”

One key to understanding these articulations – and how they implicate practical matters such as the serialization format of a data exchange – is to see how the lines and arrows in a diagram demarcate lines of authority and responsibility in a complex system of exchanges, whose most important considerations are not in the details of any single exchange, but rather in the operational context in which all of them take place together, as an orchestrated set.

By “exchange” here we mean more than simply an exchange of data. Certainly, some of these transitions entail data being copied from one system to another. (As a journal article is sent as an email attachment from its authors to the issue editors.) However, exchanges also happen as data transitions in other ways. An editor who assigns to an assistant, for example, a task such as copy editing, may “move” nothing, except assign access control rights in the system (so the assistant can make changes to the copy). Yet to exchange access control is to exchange much else, namely the custody of the article, entirely or in part. (In our example, the assistant may know that some corrections are in scope while others cannot be executed without conducting another loop outward, with the author.)

As an example of the signification of such an exchange, over and above the communication of the exchange itself: a young scholar publishing an article in a leading journal, assumes and acquires thereby some of the authority and credibility (as it were by proxy) of the journal. The published article becomes a line item in a *curriculum vitae* and eventually a tenure application, and as such might be worth as much as or more than the article itself. Of course, this is of no direct interest to readers of the article, who benefit from the scholarship despite the necessarily mixed motives behind it. Indeed in principle, the combination produces a mutual benefit even if not a perfectly symmetrical one. (The journal, and its readers, get the good scholarship. The author gets a shot at tenure.) In any case, several exchanges occur following on the central one (the article's publication) at several levels: exchanges of authority and reputation, as well as notice of interests and alliance.

This kind of thing matters since it shows how these structures are built on and around incentive structures, which is to say combinations of mandates, structured choices, and agreements to cooperate that condition how these systems are built and maintained. (One such agreement to cooperate takes the form of “I will do this task if you employ me and make it part of my responsibility” – which is at one level a significant commitment. And yet some kinds of tasks, it seems, are not routinely accomplished without it.) At question is always not only in what *form* does a data exchange occur (an email, a Word document, a piece of registered mail with a signature, a spreadsheet – or an XML document, valid to a schema?) but also *by what rule is that form determined*, *who makes that rule*, and *whose interests are served* (immediate or long-term) by that rule and its rule set, both actually and apparently? (“Make it a `docx` file because that's what I know how to use.”) In our example, the young scholar may happily take on the work of formatting the bibliography, as the journal submission guidelines demand, because she understands this exercise is both valuable in itself (or she wouldn't be asked to do it, presumably), and, in a sense, the price of admission, the cover charge for the club; a demonstration and proof of her willingness and ability to play by the journal's rules and pitch in to the common effort. (It is not entirely uncommon for scholars to find bibliographies in particular as rites of passage.)

Or alternatively (to work this example further) maybe the journal discovers that it can't get good enough bibliographies from its authors no matter what it does. So the work of reformatting bibliographies is handed to an in-house assistant. The workflow for handling the article is thus articulated, at the point of the bibliography. Custody shifts, with respect to an isolable (rules-bound, thus also “typical”) chunk of data, namely that part of the article (the bibliography or works cited section) with its links or bindings to the rest of the system (conceived in large terms). The workflow becomes more complex and the bibliography becomes a special focus. Submitted to a more stringent set of rules, by an agent or operator who can specialize in them, the bibliography can now be enhanced in ways otherwise impossible, normalized for integration. So the bibliography of the paper is made by this effort to merge more easily with the larger bibliography of the journal or publisher's holdings, not only this bibliography but all of them. The benefits of having such a mega-bibliography grow exponentially as the number of entries grow, so it is worth building as largely as possible if you're going to the trouble at all. However, this comes at a cost: someone must pay for the expert assistance and the technical stack to support it. Someone must learn how to do it. In the real world, someone has to volunteer to take on this responsibility, or there must be a budget to pay someone to do it.

It is not difficult to find examples of such phenomena, which are indeed at the heart of publishing activities or more largely, of business in general. Exchange happens, we can stipulate, when a data set in some form or representation (a “document”, an “article”, a spreadsheet, some sort of formal submission on a template), shifts from one party to another, for an operation to be performed. Submitted to such a process, there may be a new record created, and/or an original record or document may be altered or amended: in any case there is a before/after relation; in a way of speaking, each discrete step can be considered an operation, function or filter.

For each of these, whether machine aided or entirely motivated and performed “by hand”, there is some investment (cost), and some reward. In return for providing its value to the operation, the function or filtering operation must be paid for. Standards-based automation pays for itself when we can make these costs linear (by factoring out costs of design and development), while the benefits remain exponential, whenever we can operationalize such functions or filter to the point that they can be automated.

Quality is defined within context

Another key is to see how, within these transmissions, the question of *quality* is both construed (defined and determined) and maintained. Again within the context of journal publishing, a (nominally) “high-quality” author submission might well take the form of a word processor or “office”document. (Whether it is Microsoft Word, Google Docs or whatever a journal editor might consider acceptable these days.) In this case, the criteria of quality are not in its formatting – how pretty is the research laid out on the page – but rather in its instrinsic properties of argument, evidence and exposition, relating it as subject matter to other subject matter. (Is it original or novel research in its field? Does it relate to the literature in its field in some other meaningful way?) As such, the entire purpose of a produced artifact such as word processor file is to represent its author's work adequately for the purposes of the journal to “publish” it, a complex process that entails among other things (and again, because criteria of quality are extrinsic to this), that the document will be translated into a new form – for example, a PDF for page display, even eventually ink on paper.

In its new form, the reformatted document has “quality” (or one might more properly say “value”) that the original document does not, and is judged accordingly – now, not only for its argument and evidence (its nominal “content”), but also for its aesthetics and accessibility (for example).

In other words, it is worth looking at the before and after states when considering appropriate criteria of evaluation. Before publication, as submitted, we might like the document to look nice on the page; but it is not the page layout by which we judge it. This matters because part of what we intend to do, indeed, is reformat it so it looks different. In other words, we fully expect that the article or work we accept for publication, will be changed in that process, if not in essence (as FRBR “work”[5]) then in representation. Yet as publishers (to say nothing of the production designer), we expect the work to look “better” or more polished (than the author could make it). The publishing enterprise is designed to support such activities through processes that work not simply by adding value but by doing so within the context of new and more stringent criteria for evaluation, changing the definition of “quality” itself.

Moreover, this shift in what might be called the *evaluation context* for determining quality is entirely the point of the workflow, the “refinement” to which the work is subjected. Significantly, this can happen irrespective of how the worflow's participants constituent – the various players in the exchanges – are more or less oblivious to it. The young scholar is rewarded by publishing the article in the leading journal. Part of the reward is that the article is now listed in indexes; it pops up in searches. Other researchers are led to the scholar's work through these links. The links were made not by the scholar who wrote the article, but by the staff who provided its metadata and aggregators who followed after to consume the refined works (journal articles, issues and volumes) made from the raw word processor documents collected from scholars in the field. The very existence of these links, and the aggregators who make them, may be unknown to the scholar whose paper is being cited. Yet they serve a purpose, and the scholar benefits from them indirectly without knowing about it.

In a way, this is to note again that the plan or design of the “machine” of a publishing enterprise is already larger and more complex than the various machines, people and processes that are embedded in it, sometimes even larger than the participants appreciate. The special opportunities of automating document workflows, where we can do this – which is to say, the opportunities and promise of information technologies to enable things that go beyond what could have been done with ink and paper or even telegraph and telephone – will have to *accommodate* these larger systemic requirements, not work against them. Or they will simply not be viable.

The good news is that we are now at a point where we know that these systems can work, and indeed work well, when their various parts are adequate to their needs and where, just as importantly, contributors to the effort know and understand something about how the system works, and why it takes the form it does. (Even if they cannot see everything all the way to the edges.) Enough documentation projects have subsisted long enough, at various and very different levels of scale and complexity, that we can be confident of what we know about this.

Evolution works by little revolutions

A technological system exists until the day it stops being used. After that day, its relics may subsist, but the system itself does not. But a system can also be renewed over time from the inside, shedding parts of itself (since a system is made of subsystems) and replacing them, as its users continue to use the system, but modify it while using it. At one level of the organizational hierarchy, a system is brought down or replaced; it comes to an end; it is switched out for another. This same activity, seen from the next level up – from the point of view of the larger system in which this subsystem works – the switch out constitutes renewal, not death. Maybe there was a day when you used Eudora or Pine for your email, and now you do no longer. Has your email system died? Or merely migrated? Depending on how we define the system: both.

Just as your system can be someone else's subsystem, all their systems can be subsystems of yours, to the extent that you rely on them to do things for you, that you do not do for yourself. The boundaries in the system are not determined so much by extrinsic factors such as the software or platform on which it runs (especially when so much runs on the cloud), as they are by agency and scopes of responsibility – who is responsible to do what – and the interfaces and functionalities that support this.

New systems do not successfully replace old systems except (ipso facto) when they answer the needs met by the old system, and this can happen in only two ways: either the new system grows out of the old system, as it were within the context of its interfaces, and therefore replaces it organically. Or the new system is engineered to replace the old system by offering the same capabilities, perhaps along with some other definitive advantages such as scalability or ease of use. Again, this might be a single process, looked at from two directions.

So we might consider the way email has replaced sending paper through the post (mail), for most routine transactions. A publisher that once collected stacks of paper manuscripts, now pulls together file sets culled from email attachments. This development happened organically, but it would not have occurred if email had not been designed and extended to serve some of the basic (or “essential”) functions of paper mail, even while it is crucially different in other respects. It is worth recalling one of Marshall McLuhan's adages, that the content of any medium is another medium.[6] It is not quite a drop-in replacement – email promised and offers new capabilities beyond what the paper post ever did – but it is capable of many of the same functions and operations.

Any disparity of perspective here tends to be not a disparity of kind but of scale, that is again, of the level of hierarchy at which the problem is viewed. To return to the journal example, to an editor as email user, for example – `editor@journal.edu` – correspondence with contributors and readers is an ongoing and essential process, which must occur for the journal to subsist, somehow or other. This volume of information (correspondence, manuscripts, edited copy, reviews, in-process transcriptions), as a kind of information matrix, is the medium out of which the medium of the journal itself (through a kind of alchemical distillation) is made. From the point of view of the journal – a fish looking at an ocean – the case of its correspondence is distinctive, unique and special to itself. (The editor does not care for anyone else's journal correspondence, nor are they expected to.) As a team, the journal staff undertakes the responsibility of supporting this exchange with the people they wish to reach. (This is what it means to produce a journal.) Does this mean they need to develop their own postal service or messaging platform? No: in the real world, what they do is necessarily what their correspondents and partners in exchange (authors, readers) already do. In other words they do not and indeed cannot invent something new, instead, adopting as an externality a shared platform or system (the post, or email on the Internet, or a package delivery provider) already available (an externality) and indeed designed and engineered as a system, working at a higher scale, for a more general purpose than to join this journal with its authors and readers (namely to enable any such journal, and many others as well, and many activities and enterprises beyond journals). From this point of view, this journal's particular problem (as a "user", we might say) becomes only another instance of a more generalized problem – not maintaining a correspondence with readers and writers, but only an email system (or, before email existed, a postal system) among others.

It might be an interesting debate to discuss whether and in what respects the journals we produce today, with the support of electronic communications such as email and file exchange over the Internet, are "better" or "worse" than journals once produced on platforms we have long ago migrated away from. (A science of workflow might interestingly also be an archaeology of workflow.) Certainly the volume and rapidity of information exchange today is greater by orders of magnitude. There may also be shifts in who is able or permitted to participate, and for what presumed as well as actual purposes. However, we do not now prefer email, or our digital platforms of choice, to paper and postage, because they enable better work, so much as because the *scale* at which we now work – the number of partner exchanges we have, of what quality, and of what kinds of information codified in what forms – would simply not be achievable (much less sustainable) without the capabilities of the digital machine for information storage and manipulation.

We have seen a similar migration in the progression in "camera ready copy" for production of both print, and print surrogates, from literal image files, through printer instruction sets (such as PostScript ™) to today's PDF transmissions. All of these systems had to be engineered, but almost no one who adopted them paid much attention to the engineering itself. Each successive system merely met the need better than its predecessor. Today we have something far superior to what was ever possible without the networked exchange platform we now have (the Internet) and all the standards developed to support it. But no one exactly noticed much as our means of sending "pages" improved – as one subsystem replaced another. It simply happened.

Of course, it did not simply happen by itself, and the emergence of superior means (or at least, more capable means at scale) for maintaining business correspondence (email) or producing materials for the eyes of readers (camera-ready copy) reflected significant efforts by their developers and early proponents. The efforts were not made by the eventual users, however. Similarly, our users should not have to design their own technical solutions. A new platform emerges because we work at several levels of the system at once – and because we exploit emerging opportunities.

If our aim is to provide any "downstream user" with the kinds of capability and leverage one gets from external control, specification and testability of the kinds of regularities – at multiple levels of "semantics" – that can be usefully discovered or introduced into our data. We do this by working at different layers, providing users and indeed application developers not with solutions, but with the foundations and technical infrastructure on which their solutions can be constructed; but once that has been done, the solutions themselves "just work". In other words, what from one point of view, is an engineered solution, must be from another, just a better way to do the same thing.

Yet in a world where processes are already well defined and described, this is a challenge, since application design must in some way come first – at least insofar as engineers must build to specific problems, not just general ones, to motivate the efforts. Solutions will emerge – someone will put effort into building them – if the base works well enough to enable and streamline these efforts. With a combination of good data, and good data description, we believe this is possible and necessary in a domain as complex and semantically rich as this one.

How do we know we need (something like) OSCAL?

As pointed out in a 2019 paper for this conference, many or most of the functional requirements for data capture and relation (linking) that we face in systems security could be achieved, with only a bit of stretch, by extant markup technologies. The reason we need OSCAL is not because existing technologies including DITA, ISO/NISO STS, HTML microformats, XBRL, or even TEI, not capable of representing the data adequately (just to mention a few reasonable candidates[7]). Indeed if we assume that any of these has such capability, the question becomes why documents relating to systems security, and their exchange, remain firmly locked into proprietary technologies of production such as word processors and spreadsheets, given their well-understood limitations for systems working at scales beyond the enterprise. Why, in other words, has this migration not long ago happened?

One answer to this question is in plain view in the form of a common office document feature, namely templates, and the simple fact that an Office document (whether Microsoft Word or Excel, or a similar application on or off the web), for all its limitations, is the most flexible, powerful and accessible tool available (to one definition of “accessible”) to a security professional, for data modeling. And data modeling – the definition, collection, management and deployment of structured, semantic data – is indeed at the core of their work. Available encoding standards all assume one thing: that the schema that adequately describes the document for its intended application, can be known ahead of time, indeed is not only known, but anticipated and accounted for by the general-purpose schema in question.

But every new structured document and every spreadsheet implies a model, and usually one (assuming a good designer) whose outlines are readily discernible to an informed reader. Indeed frequently, documents in use must conform to a “type” and follow rules for that type – evidence of the model again – with templates used as one (not the only) criterion for measuring conformance to the type (considered informally). Now, these new models are not made just for the enjoyment of it (although that could play a factor). On the contrary, we invest the effort because we can see the benefits (in higher quality, better and more throughput – that is, data processing capacity and capability) of doing so. The rules are not an impediment but a track to follow.

Many data professionals understand the shortcoming of office documents (considered as a genre) for truly widespread data, secure data exchange. But even they have no choice but to use them, since document templates are also what their own downstream users can use – while XML application stacks and libraries of stylesheets (or even the functional equivalent for JSON) are not.

Indeed it might be said that the essence of our problem is to make it possible for data professionals to do more than encode their information optimally for exchange (only) with their *immediate* partners – however necessary this is, and great the benefits of doing so. A workable standard for active information exchange is a *sine qua non*, but beyond it is another essential, since the requirements for exchange themselves in this domain are so local and so particular to processes, and defined and mandated at several levels at once.

This meant that whatever language we adopted or developed to address the needs, its own extensibility model would be crucial. (This is not an unfamiliar problem to the designers of documentary encoding standards.) Ideally, extensibility features would be free to users in the sense that no work in a schema or formal layer should be necessary for local applications to define and then enforce their own semantics (for example, by offering features enabling extension by restriction). But mechanisms for them to introduce such enforcement are also essential. We found a solution to this issue in our Metaschema technology (as described in the 2019 paper).

From this point of view, OSCAL (the Open Security Controls Assessment Language) should be regarded not as a solution, so much as a process for developing solutions. This process is technical and entails the definition and specification of information models as means to and end (successful, meaningful data exchange). But it is also very granular, and happens “on the ground”. Like any documentary standard, OSCAL may be able to provide for 80 % of what is needed for its "normal" cases, and what constitutes a normal case for it, as all technologies, will be a direct reflection of its capabilities. What is most important, however, is how OSCAL permits its users to deal with their 20 %.

It is not the aim of this paper to describe in any detail the OSCAL models or how they address requirements: a certain amount of background information on the project might be a prerequisite for much of what follows. It is too early to say whether and to what extent any adoption of OSCAL may change or improve the actual practice of security assurance. But it is not too early to reflect further on the challenges that could impede its success. And if the basic premise of OSCAL, like other markup technologies, is in declarative markup and descriptive encoding, it seems necessary to consider what problems or issues we might watch out for.

The OSCAL Approach

This paper is for two audiences at once: information technologists who specialize in open data formats and the standards that sustain them; and systems and information security professionals who bring an understanding of the requirements of their domain, not necessarily deeply informed of available approaches or solutions (in the form of available technologies), but who bring an interest to this topic because they know or sense enough about it, to understand the significance of their impact. If anything, what makes our project interesting to the first of these audiences, is our relation with the second. Briefly, we are hoping to change the practice of systems security and its documentation, by presenting its stakeholders and practitioners with better ways of doing things. While we have some ideas (or we would not be making the attempt) of what these better ways look like, we must assume, however (quality is defined in context) that they are in a better position than we are to know what "better" will be. This means our primary challenge is to listen, with the goal of empowering them (users, stakeholders, the community, the market) to do what we would do for them, if we knew what they know.

In view of this, it may be worth considering briefly what OSCAL (the Open Security Controls Assessment Language) is, and is not. OSCAL is:

- A set of related and interlocking data models
- A data description language for a domain, and thus defined by that domain: systems security assurance and related documentary activities, as defined by IT (information technology) practice and statute.

As described in Piez 2019, these models are defined by a schema back end or *Metaschema* technology that permits us to provide support for these models in multiple syntaxes, specifically (to date) XML, JSON and YAML syntax.[8] This support takes the form of schemas for validation, conversion utilities and much else (documentation, starter stylesheets for designing representations, code generation, etc.). Together, these offer the platform for a stack of capabilities for data description, application and interchange. In this it is analogous to many other standard or common encoding technologies (XML-based and not), as they address their respective domains.

What then is OSCAL *not*?

- It is not a markup language.

  While OSCAL has an XML expression, and is designed for use in and with markup-based systems, it is also not a substitute for DITA, JATS/BITS, NISO STS, HTML[9] or any other extant markup technology, which are considered to be (from the OSCAL perspective) not alternatives (since they do not address the same set of functional requirements for data representation), but rather as complementary technologies and (as such) exploitable assets.
- OSCAL is also not an attempt to engineer a workflow or "solution" to the problem of data management in this complex domain. Rather, it is intended to provide the foundation or groundwork for the development of workflows and solutions.

  Again, existing workflows might be regarded as externalized assets and as opportunities, insofar as to the extent they can be "OSCALized" (enabled with and by OSCAL), they can work together with other systems more seamlessly, acquiring new capabilities via network effects.

So what about those functional requirements? The core concept is close to that of standards-based markup languages, albeit scoped within the particular domain of a specialized information set: we wish to enable better system security and security assurance by providing a foundation for rich (semantic) data exchange among partners and organizations. If we succeed, we will *lower the costs to organizations and users* of participating in such exchanges, by helping to *define and apply the rules* that enable it.

So how are the requirements we are addressing unlike requirements for publishing systems?

- When publishing for an audience of no more than three or four parties, requirements for production values are different, and economies of scale in production will not benefit in the same ways. Return on investment still comes from economies of scale, but not to the same (exponential) degree. At this time, achieving superior results, arguably, is as important as lowering costs. (This is not always true in publishing, which has been subject to economic stresses for much longer.)
- With respect to the data sets themselves, the granularity of description they require is (as compared to many applications of markup languages) relatively *rough*. This is reflected in the fact that OSCAL applications do not need much functionality at the word and phrase level – the data models is seeks to capture generally do not require it – and that when it comes to discursive contents ("prose"), it does not need much beyond some inline formatting plus a generalized insertion or transclusion mechanism (somewhat analogous to DITA **key/keyref**) working at the phrase level. Since these together can be accommodated using a near-subset of HTML or Markdown, the information can also (with some compromise) be constrained to representations that fit well within the limitations of JSON or similar object notations. In the terms I used in my Balisage 2018 paper Piez 2018, the data is higher on the "semantic stair".

  The flip side of this is that at a higher level of granularity – groupings of prose and structured data – the requirements for what might be called "hypertext" are comparatively intricate . Documents and their parts and components present complex interlinkings, both to one another and to similar or different parts of similar or different but related documents. For example, System Assessment Plans must make targeted references from their parts, to parts of system descriptions given in System Security Plans. Those references are "semantic" in the sense that they must be distinguished by type according to intended use and the kind of relations they encode; and when the links break, the documents break.
- Above all, OSCAL's users are different from the users of publishing systems or even from operators of documentary-production workflows.

  Most importantly, OSCAL's users will not only be CMEs (content matter experts) who use OSCAL-based systems to acquire and represent data to do their security assessment jobs, but also developers of systems and software to use it. While we do not expect that most OSCAL data will be published widely (the exception being canonical documentation such as the catalogs and baselines to which other OSCAL documents refer), we do expect it to be useful within organizations and between partners, in multiple unforeseeable ways. This means that developers need to be able to build to it.

  In order to win their support as well as maximize the chances for their success, we do not wish to constrain, any more than absolutely necessary, those developers with any encumbrances with respect to formats, software platform(s), or technical dependencies in general. Because we expect and rely on them to take us places we cannot go, we must trust them to use the means that seem most appropriate to them.

  In our experience, most dev/ops professionals become literate in multiple different formats for information interchange. However, it is also important to meet them where they are, and to enable them to use tools they know (while opening opportunities to use tools they do not yet know).
- This means that while we are free to define, demonstrate and promote models that enable functionality to be delivered, we are not free to stipulate that only XML may be used. Today, it is either JSON and other formats (including but not limited to XML), or JSON only.

  And indeed, since the information in our domain is not only documentary and not only suited to XML, having the capability to work in either format is a huge advantage. And designing from the start to be able to support and address either, also positions us over the longer term – as adding support for yet more alternatives (YAML, for example) becomes easier to do.

Notwithstanding these differences, both the complexity of the requirements, and the "documentary" and "fractal" nature of the data sets themselves – they exhibit regularities, but they are not entirely regular – necessitate the layered approach to systems design. As stated at the outset, we believe that a layered system that relies on declarative, descriptive encoding of data structures, with a separation from both underlying platforms (operating systems, storage media etc.) and from application logic, is the only practical approach to dealing with information management this complex. Indeed, the problem presented by systems security (planning, analysis, implementation, documentation, assessment) – whether considered as "micropublishing" or not – might be described as similar to the problem of designing robust, sustainable (platform-independent) publishing systems, except "on steroids": perhaps an order of magnitude more complex. The only way we have of managing that complexity is to factor it out into separate interrelated sets of requirements, which can be addressed separately as well as together.

If we have built and operated successful systems on that basis, the next question becomes what that experience teaches us.

Applying the lessons: conclusions and expectations

To state that there is a way forward is not to say that it will be easy. Considering the lessons of XML in publishing, with its challenges, can help us reflect on the challenges we also face:

- **Your system is someone else's subsystem**

  This is even more true in the realm of RMF-based security assurance activities than it is in mainstream publishing. By design, an OSCAL document produced in and for one system, will be used, worked and integrated within another. An OSCAL system component description, for example, can be encoded once in one resource, then referenced by many systems plans that integrate that component. These documents will all be composed, produced and shared in different organizations (such as when the component in question is developed by one vendor and then used as a platform by another); and the links must hold together.

  One of the keys to scaling within the publishing domain is that an article, monograph or any published “artifact” is conceived as, in principle, a self-contained and self-sufficient entity, which can be written and produced for publication separately from others of its kind. Of course (as discussed above), this self-containment is partial and qualified, as much of the work of publishing is the integration of such an entity with other such entities within a larger structure – articles are integrated into a journal and even monographs are anchored into larger infrastructures for purposes of marketing, distribution, cataloging and so on. This is achieved through the application of two principles: (1) distinguishing “kinds” or classes of publication that can be treated alike within larger systems; and (2) associating metadata with each publication that can distinguish the instance among the members of the class.

  A primary goal of OSCAL is to begin to do this, and to provide a foundation for continuing to do so, within the domain of systems security documentation. Of course, the idea of kinds or classes of documents brings us to declarative markup.

- **Declarative markup adds value**

  The layering that is characteristic of systems based on declarative principles has been well explored at this conference and elsewhere. This layering enables separation of concerns between the production of data content (information sets), and its subsequent management, processing, rendering (presentation) and downstream application. When applied to publishing, this principle works.

  Again, however, the fact that there is a principle we can follow, does not make the design problem easy. In this case, distinguishing meaningful classes of information according to their nature, purposes and uses, depends on a clear and articulated sense of both commonalities across, and boundaries between different information sets, as well as their complex relations. Shared documentary structures that reappear throughout OSCAL's models reflect these commonalities; rules on their use reflect the boundaries. But both the shape of those structures and the rules applied to them, must make sense in terms of the data set as the practitioner sees it.

  And because our view of this is partial and evolving, we have also made efforts to enable the modeling to be agile and flexible and adaptive, most especially in the back end (Metaschema) technology we have developed to enable modeling across the gap between XML and object notations (see Piez 2019).

- **Data acquisition is hard**

  If only because it is so challenging in other domains, we should be ready to place special focus on developing means to convert relevant data sets into well-structured, well-described OSCAL.

  Multiple methods and approaches could be explored, using an “all of the above” strategy: structured editors; forms interfaces; semi-structured resources such as wikis or “issue” (ticket) systems; office document conversion pathways. Different methods or strategies may be appropriate for different parts of the system.

- **Activities are supported by incentive structures**

  The incentive structures within this domain are very different from publishing, and positive incentives are arguably scarce. Historically, activity related to systems security (beyond the functional minimum) has too often been a low priority, a kind of optional insurance policy for cases where the implicit security model of “trust your neighbors and leave the door unlocked” has failed. Similarly, the second-order benefits of well-documented systems security (exposure of latent issues; traceability; assurance; contingency planning) have been considered at best as “nice to haves”. Without the heavy hand of regulation, security typically gets little if any attention from designers or developers until after a system is implemented and stakeholders are happy with its functionality and performance.

  In order for RMF-based activities to be fruitful, we need to keep incentive structures in mind, and look for opportunities to provide positive as well as negative incentives. To be sure, impediments must also be removed for positive incentives to come into play – thus for the OSCAL project we have done our best to address non-negotiable operational requirements (such as “XML versus JSON”) in ways that make it possible for deployed systems to actually start to realize benefits in data interchange.

Noteworthy positive incentives should include, first and foremost, improved capabilities: more and better risk management at less expense. The more substantial positive incentives might take the form of better and more secure systems – that is, not only the documentation, but the systems themselves will be more secure, while also easier to develop, test, reuse and adapt in a “security first” mindset.

Additionally there are important secondary incentives, such as a more efficient use of time invested in assessment when the overhead of manual operations is reduced. Given that there is always more to assess, it is difficult to imagine how security assessments themselves can become cheaper. But with the aid of machining, and given better and more consistent, more easily consumed artifacts to represent the subjects of their assessment, one could expect assessments in general to be better, even to the point that “light touch” assessments (of well-documented, well-vetted systems) might be deemed to be adequate.

There are perhaps some further benefits of automation and automatability that might become incentives, to the extent that it can be recognized how achievable they are given appropriate investment. Much of the design of OSCAL is intended so that the considerable efforts of authors at lower layers – people who define and publish catalogs and baselines – can be better leveraged and exploited by the consumers of their information whether that be planners, assessors or others responsible for defining policy. This could become an incentive were it possible to monetize or otherwise feed back that benefit. (Pay a license to use an especially good baseline?) More likely, it becomes an incentive to the extent that such use and reuse of one's catalog (or baseline) is itself considered a criterion for success.

Finally, it is worth stressing that a significant factor in getting work done – to say nothing of good work – is inevitably in the imponderable aspects of pride and satisfaction with good work that good workers cultivate. It may be possible, by developing good tools, to build in “micro incentives” in the form of opportunities for good work, which serves as its own reward for those who see how consistency, transparency and integrity of the data they produce can contribute to the soundness of the system as a whole.

- **Quality is defined within context**

  With respect to security-related activities in general, or even RMF-based activities in particular, because operational context is hard to define and open-ended, no single solution will be a comprehensive solution. In the publishing domain, we have learned that the high degree of requirements for local adaptation and customizability has been critical to the success of the standard encoding formats. This is likely to be even more true for us.

  Tolerance of variation – and recognition of variation as a source of information – is an important characteristic of these systems. It being difficult to distinguish in general where variation is meaningful – where it is signal, and where it is noise – these systems need to be well defined, well managed and transparent, but also flexible and adaptable, with extension mechanisms that permit local adaptation without unnecessary “forking”.

  Although it is outside the scope of this paper, the design of OSCAL's schemas and validation infrastructure permits addressing this set of issues in ways already familiar to designers and users of publishing systems: namely, by deploying not a single “one size fits all” validation regimen, but rather by supporting a layered or tiered approach, “mix and match”. This permits organizations to define their own rules and rules sets and gain leverage over their own data, for their own (and partners’) processes, even while they also conform to a more general set of rules shared by everyone.

- **Evolution works by little revolutions**

  Everything we have learned about the application of information technologies to publishing suggests that in this domain as well, progress towards better practices and more capable systems will be incremental before it is systemic. With a view to this likelihood, OSCAL aims to offer early rewards for users who can adopt it for solving problems, whatever those problems are, without imposing a requirement for any top-down overhaul. Even when OSCAL is never used at the core of a documentary system, it might be useful at its interfaces. And if it is useful at the edges of any system, it will eventually be useful at the core of others.

  Yet experience also suggests that developers and stakeholders must “get it”, for progress to happen at all. There is no substitute for understanding, and thus for data transparency to the extent practical and possible. A commitment to open, non-proprietary declarative encoding – even within a secure operational context – is crucial, if only because a monoculture is not secure. But it is not only because of its long-term security, that the system must be open; it is because its success will depend above all on who understands it and how well, and how well they can adopt it for appropriate and intended use.

Within this context, one final principle might be recognized: *the platform is not the capability*. This might, indeed, be considered to be a core insight, in the sense that the entire enabling paradox of declarative markup is based on it – by defining the encoding in a way *independent of and abstracted from* its local application, we enable applications not only locally (any application must be local) but in general. While a technical platform (considered in the broadest sense) is necessary for a technical capability, this practical dependency is a reflection of the fact that the logical dependency is the other way – unless it enables a meaningful capability, a platform or technical means remains inert and ineffective. A platform that offers no useful capability, will soon be abandoned. Conversely, while it is necessary for developing and demonstrating a capability, the very fact that a requirement can be described without commitment to a platform, is an indication that no platform or technical solution is a *sine qua non*. The different strengths of different technologies (whether XML/XDM, Javascript/JSON, comma-delimited values exported and imported into spreadsheets, or anything else) give them comparative advantages – and these can be exploited. Thus a platform that is developed to enable capabilities we already understand – and already have the means to accomplish – can also be a springboard.

References

[declarative-bibliography] "Declarative Markup: An Annotated Bibliography" See https://markupdeclaration.org/resources/bibliography.html.

[rmf2018] Joint Task Force Transformation Initiative, "Risk management framework for information systems and organizations: a system life cycle approach for security and privacy," National Institute of Standards and Technology, Gaithersburg, MD, NIST SP 800-37r2, Dec. 2018. `NIST.SP.800-37r2`

Lubell, Joshua. "Integrating Top-down and Bottom-up Cybersecurity Guidance using XML." Presented at Balisage: The Markup Conference 2016, Washington, DC, August 2 - 5, 2016. In Proceedings of Balisage: The Markup Conference 2016. Balisage Series on Markup Technologies, vol. 17 (2016). `10.4242/BalisageVol17.Lubell01`

Lubell, Joshua. "Using DITA to Create Security Configuration Checklists: A Case Study." Presented at Balisage: The Markup Conference 2017, Washington, DC, August 1 - 4, 2017. In Proceedings of Balisage: The Markup Conference 2017. Balisage Series on Markup Technologies, vol. 19 (2017). `10.4242/BalisageVol19.Lubell01`.

Lubell, Joshua. "SCAP Composer: A DITA Open Toolkit Plug-in for Packaging Security Content." Presented at Balisage: The Markup Conference 2019, Washington, DC, July 30 - August 2, 2019. In Proceedings of Balisage: The Markup Conference 2019. Balisage Series on Markup Technologies, vol. 23 (2019). `10.4242/BalisageVol23.Lubell01`.

[Lubell 2020] Lubell, Joshua. "A Document-based view of the Risk Management Framework." Forthcoming at Balisage: The Markup Conference 2020.

[McLuhan 1964] McLuhan, Marshall. *Understanding Media*. 1964. Cambridge and London: The MIT Press, 1994.

[a-130] Office of Management and Budget Circular A-130, *Managing Information as a Strategic Resource,* July 2016. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/circulars/A130/a130revised.pdf.

[OSCAL on the web] "OSCAL: the Open Security Controls Assessment Language." https://pages.nist.gov/OSCAL/ (accessed Mar. 24, 2020).

[Piez 2018] Piez, Wendell. "Fractal information is." Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018. In Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies, vol. 21 (2018). https://doi.org/`10.4242/BalisageVol21.Piez01`.

[Piez 2019] Piez, Wendell. "The Open Security Controls Assessment Language (OSCAL): schema and Metaschema." In *Proceedings of Balisage: The Markup Conference 2019*. Balisage Series on Markup Technologies, vol. 23 (2019). `10.4242/BalisageVol23.Piez01`.

[Piez 2001] Piez, Wendell. "Beyond the Procedural vs Descriptive Distinction." Extreme Markup Languages 2001. Archived at http://wendellpiez.com/resources/publications/beyonddistinction.pdf

[Tillett 2004] Tillett, Barbara. "What is FRBR? A Conceptual Model for the Bibliographic Universe." Library of Congress Cataloging Distribution Service. Revised February 2004. Archived at https://www.loc.gov/cds/downloads/FRBR.PDF.

Walsh, Norman, and Bethan Tovey. "The Markup Declaration." Presented at Balisage: The Markup Conference 2018, Washington, DC, July 31 - August 3, 2018. In Proceedings of Balisage: The Markup Conference 2018. Balisage Series on Markup Technologies, vol. 21 (2018). `10.4242/BalisageVol21.Tovey01`

[xdm2017] XQuery and XPath Data Model 3.1. https://www.w3.org/TR/xpath-datamodel/

[1] SGML is "Standard Generalized Markup Language", ISO 8879:1986.

[2] LaTeX, a document processing system, is hosted at https://www.latex-project.org/.

[3] RMF is the *Risk Management Framework*, an approach to systems security management and documentation codified in NIST Special Publication (SP) 800-37. See rmf2018 and Lubell 2020.

[4] “Rheology” is a branch of physics. A science of workflow would be a branch of data science and cybernetics, applied at the level of the human organization, drawing (at least) from sociology, economics, general systems theory and operations research.

[5] FRBR is the Functional Requirements for Bibliographic Records, a model defining categories of description for bibliographical objects such as “books” and “articles” so that “the same” (book or article) can be distinguished and related systematically even across different variants or representations, including editions, translations, printings and copies. The “work” is the highest and most abstract category within FRBR. See Tillett 2004.

[6] As he writes in Understanding Media (McLuhan 1964 p. 8), “… the ‘content’ of any medium is always another medium. The content of writing is speech, just as the written word is the content of print, and print is the content of the telegraph.”

[7] DITA: the Darwin Information Typing Architecture; ISO/NISO STS: the Standards Tag Suite; for HTML microformats see (for example) schema.org; for XBRL see https://www.xbrl.org/for TEI (Text Encoding Initiative) see https://tei-c.org/.

[8] YAML is “YAML Ain't Markup Language” (web site https://yaml.org/), a notation describing an abstract data structure amenable to processing in object-oriented languages. Its data model is an enhanced superset of the JSON object model; so by aligning with the requirements of JSON, an object model is thereby also expressible in YAML.

[9] JATS is the “Journal Article Tag Suite”, an encoding standard hosted at the National Information Standards Organization (NISO), hosted at https://www.niso.org/standards-committees/jats. BITS is “Book Interchange Tag Set”, a related encoding system hosted at the National Center for Biomatics Information, National Library of Medicine (NIH/NCBI); see https://jats.nlm.nih.gov/extensions/bits/. NISO STS is “Standards Tag Suite”, a related encoding system designed specifically to support the publication and maintenance of technical standards documents: see https://www.niso.org/standards-committees/sts.

Approximate word count: 13125. 2 figures.

Balisage:* *The Markup Conference

Effect of Amplitude Mismatch on Entanglement Visibility in Photon-Pair Sources

Paulina S. Kuo
Information Technology Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, Maryland 20899, USA
paulina.kuo@nist.gov

Abstract: Entangled photons produced by parametric down-conversion effectively have two down-conversion paths. Ideally, amplitudes of the two paths are matched. We show that the entanglement visibility is, to first order, insensitive to amplitude mismatch. © 2020 The Author(s)

Entangled photon-pair sources are typically based on spontaneous parametric down-conversion (SPDC). In a type-II SPDC polarization-entangled source, the generated entangled state is

$$|\psi\rangle = \alpha|HV\rangle + \sqrt{1-\alpha^2}|VH\rangle, \tag{1}$$

where α is the amplitude ratio parameter, $|HV\rangle$ refers to horizontally polarized signal and vertically polarized idler, and $|VH\rangle$ refers to vertically polarized signal and horizontally polarized idler. The two terms in Eq. 1 correspond to two different biphotons that can be produced by two different physical paths [1], two different crystals [2], two different quasi-phasematching periods [3], or two different down-conversion paths in an aperiodically poled crystal [4, 5].

In an ideal polarization-entangled state, the amplitudes of the two terms are equal. When measuring coincidence counts in the diagonal/anti-diagonal bases, in order to get zero coincidences (required for high entanglement visibility), there must be perfect destructive interference of the $|HV\rangle$ and $|VH\rangle$ photon paths (assuming no background noise). If the amplitudes in Eq. 1 are unequal, then this destructive interference is not possible. In several of the SPDC techniques mentioned above, the amplitude ratios are fixed during fabrication and it seems that good entanglement visibility can still be obtained even if the amplitudes are unbalanced [3–5].

High entanglement visibility requires high indistinguishability between the two possible down-conversion paths. We can quantify this indistinguishability by considering Hong-Ou-Mandel (HOM) interference between the two down-conversion paths. Imagine interfering the two down-conversion paths (or two biphotons) on a beam splitter. If the paths are indistinguishable, then no coincidences will be observed between the two outputs of the beam splitter, resulting in the well-known HOM dip. The closer the HOM dip is to zero, the better the indistinguishability and the better entanglement visibility we expect.

We can calculate the depth of the HOM dip by considering the joint spectral intensity distribution of two photons exiting opposite ports of a beam splitter, which is given by [6, 7]

$$I(\omega_1,\omega_2) \propto \frac{1}{2}\left|C(\omega_1,\omega_2)e^{i(\omega_1 t_1+\omega_2 t_2)} - C(\omega_2,\omega_1)e^{i(\omega_2 t_1+\omega_1 t_2)}\right|^2, \tag{2}$$

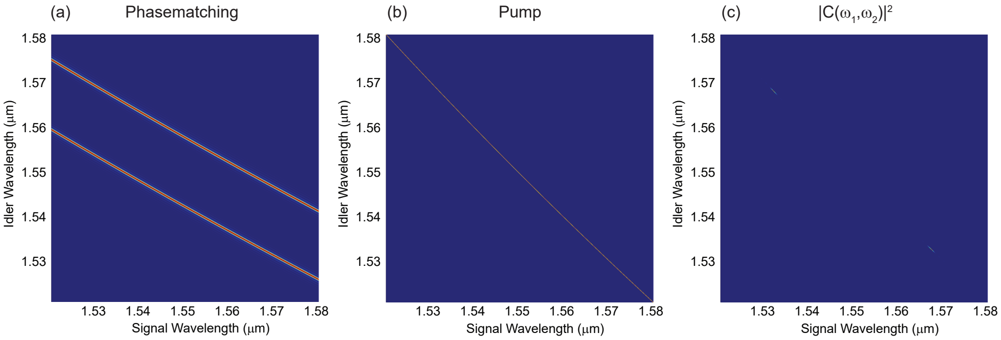


Fig. 1. The (a) phasematching and (b) pump functions are multiplied to produce (c) $|C(\omega_1,\omega_2)|^2$. The case of equal amplitude is shown.

Kuo, Paulina. "Effect of Amplitude Mismatch on Entanglement Visibility in Photon-Pair Sources." Paper presented at Integrated Photonics Research, Silicon and Nanophotonics (IPR), Montreal, CA. July 13, 2020 - July 16, 2020.

where $C(\omega_1,\omega_2)$ is the joint spectral amplitude of the two photon wave-function incident on the beam splitter, ω_1 and ω_2 are the frequency of the two photons, and t_1 and t_2 are their corresponding arrival times. The coincidence detection rate for obtaining counts in both output ports of the beam splitter, R_c, is [6]

$$R_c \propto \iint d\omega_1 d\omega_2 I(\omega_1,\omega_2). \tag{3}$$

When the photons arrive at the same time ($t_1 = t_2$) and $C(\omega_1,\omega_2)$ is symmetric (i. e., $C(\omega_1,\omega_2) = C(\omega_2,\omega_1)$, then $R_c = 0$ and we expect perfect indistinguishability of the two down-conversion paths.

As a specific example, we studied the down-conversion process presented in Ref. [5]. A domain-engineered, lithium niobate crystal is designed to simultaneously phasematch both $|HV\rangle$ and $|VH\rangle$ processes for the wavelengths 775 nm $\longrightarrow$ 1533 nm + 1568 nm. We used a narrowband pump that constrains the idler frequency, ω_2, to be related to the pump, ω_p, and signal, ω_1, frequencies by $\omega_2 = \omega_p - \omega_1$. The $C(\omega_1,\omega_2)$ function is the product of the phasematching and pump functions, as shown in Fig. 1. The phasematching function has two lines for the two simultaneous down-conversion processes.

Using Eq. 3, we calculated the coincidence rate as a function of the mismatch in amplitudes the two down-conversion paths. Figure 2 plots R_c as a function of peak ratio, which is equal to $(1-\alpha^2)/\alpha^2$. R_c is normalized to 1 as the peak ratio goes to zero. We see that when the peak ratio is near 1, R_c is near zero and is to first order independent of the peak ratio. This indicates that the two down-conversion paths are nearly indistinguishable and that entanglement visibility is not sensitive to small mismatches in amplitudes of the two down-conversion processes.

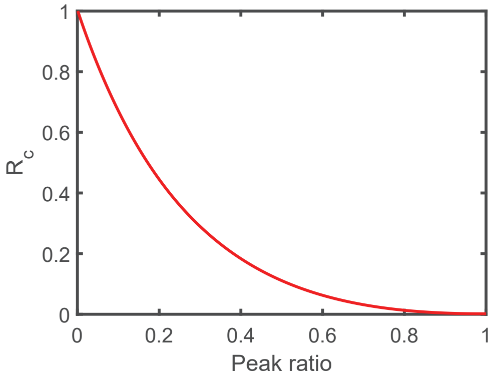


Fig. 2. Coincidence rate as a function of peak ratio.

In conclusion, we quantify the indistinguishability between two down-conversion paths by considering an analogy to HOM interference. We observed that mismatches in amplitudes between the $|HV\rangle$ and $|VH\rangle$ states do not significantly affect R_c and in turn, do not significantly degrade the polarization entanglement visibility. This observation explains how SPDC sources whose properties are fixed during fabrication can still have high polarization entanglement visibility even in the presence of fabrication imperfections.

References

1. P. G. Evans, R. S. and Bennink, W. P. Grice, T. S. Humble, and J. Schaake, “Bright Source of Spectrally Uncorrelated Polarization-Entangled Photons with Nearly Single-Mode Emission,” Phys. Rev. Lett. **105**, 253601 (2010).
2. P. Trojek and H, Weinfurter, “Collinear source of polarization-entangled photon pairs at nondegenerate wavelengths,” Appl. Phys. Lett. **92**, 211103 (2008).
3. W. Ueno, F. Kaneda, H. Suzuki, S. Nagano, A. Syouji, R. Shimizu, K. Suizu, and K. Edamatsu, “Entangled photon generation in two-period quasi-phase-matched parametric down-conversion,” Opt. Express **20**, 5508–5517, (2012).
4. C.-W. Sun, S.-H. Wu, J.-C. Duan, J.-W. Zhou, J.-L. Xia, P. Xu, Z. Xie, Y.-X. Gong, and S.-N. Zhu, “Compact polarization-entangled photon-pair source based on a dual-periodically-poled Ti:LiNbO$_3$ waveguide,” Opt. Lett. **44**, 5598–5601, (2019).
5. P. S. Kuo, V. B. Verma, and S. W. Nam, “Demonstration of a polarization-entangled photon-pair source based on phase-modulated PPLN,” OSA Continuum **3**, 295–304 (2020).
6. T. Gerrits, F. Marsili, V. B. Verma, L. K. Shalm, M. Shaw, R. P. Mirin, and S. W. Nam, “Spectral correlation measurements at the Hong-Ou-Mandel interference dip,” Phys. Rev. A **91**, 013830 (2015).
7. P. S. Kuo, T. Gerrits, V. B. Verma, and S. W. Nam, “Spectral correlation and interference in nondegenerate photon pairs at telecom wavelengths,” Opt. Lett. **41**, 5074–5077 (2016).

National Institute of Standards and Technology
U.S. Department of Commerce

Effect of Amplitude Mismatch on Entanglement Visibility in Photon-Pair Sources

INFORMATION TECHNOLOGY LABORATORY

Paulina S. Kuo (paulina.kuo@nist.gov)
Information Technology Laboratory, NIST, Gaithersburg, MD 20899

Introduction

Entangled photon pairs are a key resource for quantum information systems. Entangled photon pairs are typically produced by spontaneous parametric down-conversion (SPDC). High entanglement visibilities are obtained for many SPDC sources even when there are non-ideal fabrication or experimental conditions. In this work, we consider how amplitude mismatch in the SPDC process can still lead to high entanglement visibility.

Entangled Photon Pair Generation

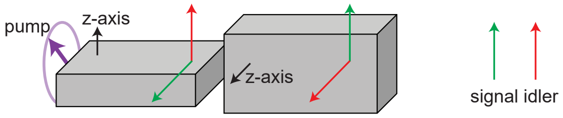
- In a Type-II, polarization-entangled SPDC source, the generated wavefunction is

$$|\psi\rangle = \alpha|HV\rangle + \sqrt{1-\alpha^2}\,|VH\rangle \qquad (1)$$

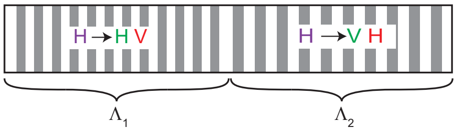
where $|HV\rangle$ refers to horizontally (H) polarized signal and vertically (V) polarized idler, $|VH\rangle$ refers to vertically polarized signal and horizontally polarized idler, and α is the amplitude parameter.

- When $\alpha^2 \neq 1/2$, the amplitudes are mismatched
- A state such as shown in (1) can be produced many ways including:

Two co-rotated SPDC crystals



Two quasi-phasematching (QPM) periods

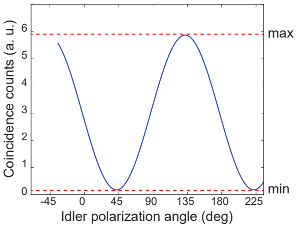


or using dual-periodically-poled or domain-engineered SPDC; see

C. Sun, et al., Opt. Lett. **44**, 5598 (2019)
P. S. Kuo, et al., OSA Continuum **3**, 295 (2020)

- The two terms in Eq. (1) correspond to two biphotons that take two different down-conversion paths (through two crystals, two QPM periods, etc.)
- In several schemes to generate Eq. (1), α is fixed during fabrication. In these cases, experiments have shown good entanglement visibility even when $\alpha^2 \neq 1/2$.

Entanglement Visibility and the HOM Interference Dip



- In a typical polarization entanglement measurement, coincidence counts are recorded while rotating the idler polarization and holding the signal polarization fixed
- The visibility, V, is calculated by

$$V = \frac{\max - \min}{\max + \min} \qquad (2)$$

- Maximum visibility (V = 1) is obtained when min = 0
- Neglecting noise, in Type-II SPDC, when the polarization analyzers for signal and idler are both H, there are no coincidences (min = 0)
- To get min = 0 when the signal is diagonally polarized, there must be perfect destructive interference between the two idler down-conversion paths, akin to perfect Hong-Ou-Mandel (HOM) interference

→ Examine the HOM interference dip to understand visibility

Spectrally Resolved HOM Interference

- HOM interference is a measure of indistinguishability. The HOM dip allows quantification of the effect of mismatch
- The joint spectral intensity distribution of two photons exiting opposite ports of a beam splitter is[1]

$$I(\omega_1, \omega_2) \propto \frac{1}{2}\left|C(\omega_1, \omega_2)\exp[i(\omega_1 t_1 + \omega_2 t_2)] - C(\omega_2, \omega_1)\exp[i(\omega_2 t_1 + \omega_1 t_2)]\right|^2 \qquad (3)$$

$C(\omega_1, \omega_2)$ = Joint spectral amplitude of the two-photon wave-function incident on the beam splitter
ω_1, ω_2 = Frequencies of the two photons
t_1, t_2 = Corresponding arrival times of the two photons

- The coincidence detection rate, R_c, for obtaining counts in both output ports of the beam splitter is[1]

$$R_c \propto \iint I(\omega_1, \omega_2)\, d\omega_1 d\omega_2 \qquad (4)$$

[1] T. Gerrits, et al., Phys. Rev. Lett. **91**, 013830 (2015)

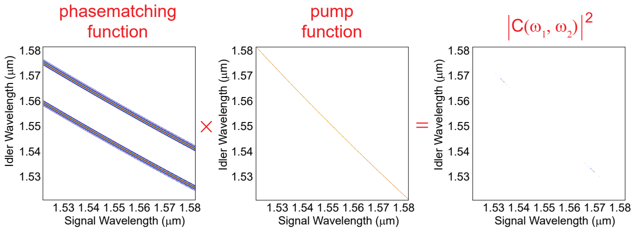
- Plotting R_c vs. $t_1 - t_2$ traces out the HOM dip

Joint Spectral Amplitude

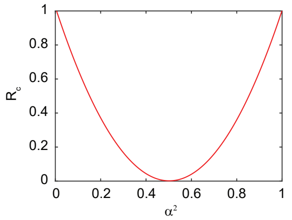
- We modeled Type-II SPDC in periodically poled lithium niobate having two QPM periods and a narrowband pump

775 nm ⟶ 1533 nm + 1568 nm

- The joint spectral amplitude, $C(\omega_1, \omega_2)$, is the product of the phase-matching and pump distributions



HOM Dip when $\alpha^2 \neq 1/2$



- Using Eq. (3) and (4), we calculated the minimum of the HOM dip (given by R_c with t_1=t_2) when α^2 is varied
- At $\alpha^2 = 1/2$, there is perfect destructive HOM interference and $R_c = 0$
- Near $\alpha^2 = 1/2$, R_c varies quadratically with α^2, which means that to first order, the HOM dip minimum is independent of changes in α^2, i.e. amplitude mismatch

→ By the previous arguments, the entanglement visibility is to first order independent of the amplitude mismatch

Conclusions

- We quantify the indistinguishability between the two down-conversion paths by considering the analogy to HOM interference
- The HOM dip and the entanglement visibility are robust to small amounts of amplitude mismatch (deviations from equal amplitudes)
- Small mismatches in amplitudes, such as those caused by non-ideal fabrication or experimental conditions, do not seriously degrade the entanglement visibility

Kuo, Paulina. "Effect of Amplitude Mismatch on Entanglement Visibility in Photon-Pair Sources." Paper presented at Integrated Photonics Research, Silicon and Nanophotonics (IPR), Montreal, CA. July 13, 2020 - July 16, 2020.

Towards inter-operable enterprise systems – graph-based validation of a context-driven approach for message profiling

Elena Jelisic[1], Nenad Ivezic[2], Boonserm Kulvatunyou[2], Scott Nieman[3], Hakju Oh[2], Sladjan Babarogic[1], Zoran Marjanovic[1]

[1]Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia
{elena.jelisic,sladjan.babarogic,zoran.marjanovic}@fon.bg.ac.rs
[2]National Institute of Standards and Technology, Gaithersburg, MD, USA
{nivezic,serm,hakju.oh}@nist.gov
[3] Land O'Lakes, Shoreview, MN, USA
stnieman@landolakes.com

Abstract. Providing the business context has a potential to become a powerful mechanism for the interoperable usage and efficient maintenance of message standards. In the literature, there are multiple techniques for its representation and application. Industry use cases have identified multiple issues that come with currently used techniques, which can represent that context. Initial assessment of a logic-based technique for doing so has been conducted and showed that some of the identified issues can be resolved. This paper presents plan for validation of the logic-based technique where proposed algorithms will be assessed in realistic integration scenarios. The paper gives details of the validation steps, goals that need to be met, and indicates issues that guide our future research plans.

Keywords: Business context, Services integration, Enterprise interoperability.

1 Introduction

Distributed and autonomous services are increasingly becoming components of the manufacturing and logistics enterprises. To achieve efficient integration of these services, message standards are needed. Messages (i.e., payload or business document specifications) define types of transaction-related information that need to be exchanged among different services. Messages are necessarily standardized because of the wide variety of inputs to those services. Inputs come from numerous business sectors, business processes, business contexts, and business representatives.

However, traditional message standards have been only partially successful. We have identified four major issues inhibiting the adoption of messaging standards. First, "out-of-the-box" message standards are necessarily generic, making them difficult to adapt to specific use cases. Second, the standards are typically documented in text form to facilitate human interpretation and do not have a computational form that can be processed by validation tools to assure integration and interoperability. Third, over time

Ivezic, Nenad; Jelisic, Elena; Kulvatunyou, Boonserm; Nieman, Scott; Oh, Hakju; Marjanovic, Zoran. "Towards inter-operable enterprise systems - graph-based validation of a context-driven approach for message profiling." Paper presented at APMS 2020 International Conference Advances in Production Management Systems (APMS 2020), Novi Sad, RS. August 30, 2020 - September 03, 2020.

the standards become large supersets of duplicative data elements contributed by various industries making the standards difficult to implement. And, fourth, as more elements are added, detailed refinement or profiling of a message standard is necessary to recapture the original business intent and business context.

On-going work by the Open Application Group [1], industry, and National Institute of Standards and Technology (NIST) [2] continues to explore the concept of using business context to improve 1) the enablement of new syntax-neutral message standards and 2) the resolution of the identified issues [3, 4, 5]. Some of the currently implemented context schemas and management approaches, however, have proved to be inadequate [6]. On the other hand, some newly proposed, promising approaches are introduced [7] but must be tested with realistic industry situations to establish their capability and feasibility. Yet, such testing is resource-intensive and requires commitment and engagement of industry and standardization communities.

We have started to validate a newly proposed approach using existing scenarios already contributed to OAG-i by industry stakeholders [1, 6]. Further industry adoption of the proposed context management approach depends on its capability to address all real-world use cases. Accordingly, the theme for this paper is our proposed approach to validate a promising, context-management technique for cross-industry use cases. Since the outcome of our approach depends on the validation process, this paper provides a detailed description of that process. That description identifies the goals and steps needed to achieve those goals. To the best of our knowledge, there are no research papers that deal with validation of a proposed, business-context-management approach.

In the following, Section 2 introduces background information about important concepts that are used in the paper, Section 3 describes the validation process through steps needed to achieve identified validation goals, Section 4 describes validation domain, and, finally, and Section 5 discusses the expected results and future research steps.

2 Background

Message standards are expressed in terms of message components. Message components are developed and offered as parts of a message standard suite where they are used to form multiple types of messages. An important, international standard, Core Component Technical Specification (CCTS), has been proposed to enable uniform and consistent development of message components and message standards. The goal of CCTS is to advance interoperability among applications and services [8].

A key concept in CCTS is *business context*, which is used to describe the business intent of using a message component or message profile, which is a subset of a corresponding message standard. In our view, capturing that intent is essential to finding and reusing the right message components and profiles, which will enable interoperability of corresponding data and services. Following CCTS, a specific business context is represented as a set of context values that are associated with their corresponding context categories (e.g., business process party profiles, geo-political location, tasks with a business process, resources available, and industry).

While the business context can indeed identify situations in which some components will be reused, we found this to be insufficient for interoperability. This is because semantic interoperability must also consider what happens in the background (e.g., transformation of the data into the database). Hence, in this paper, we assume that business context must also provide semantic definitions and validation rules that may be triggered when choosing the components in the business context. Providing those definitions and rules requires business-context knowledge.

Business-context knowledge can be represented using multiple, modeling techniques including graphical, object-oriented, logic-based, and ontology-based. For the purpose of this paper, we will employ a logic-based modeling technique - Enhanced UN/CEFACT's Context Model (E-UCM) [7]. E-UCM represents business-context knowledge using decentralized, directed, acyclic graphs. Each business-context category has at least one such graph that represents a list of possible values (or nodes) in the corresponding category. Then, a specific business context for a specific, information-exchange situation corresponds to a collection of nodes belonging to specific categories in the E-UCM graph.

E-UCM methodology for creating these graphs has two main parts. First, there is an infrastructure for (a) business-context presentation and expression and (b) contextualization of existing message profiles. Contextualization is supported by algorithms that can detect message components that are either relevant or irrelevant for a *Requested business context*. Second, there is a set of algorithms that can be used to generate the structure of new message profiles using existing message components that are relevant for *Requested business context* of a new message profile.

To determine a business context for a specific, information-exchange situation, we identify a collection of graph nodes that make up that business context. A list of specific nodes can be identified from E-UCM graph using E-UCM-provided expressions that are built using available operators and predicates. Every message component and profile has an associated E-UCM expression that captures its usage intent. This expression is called *Assigned business context*.

Along with the *Assigned business context*, E-UCM introduces terms *Overall* and *Effective business context* and provides directives for calculation of these terms for all situations and components, ultimately resulting in *Effective business context* for each component. The *Overall business context* can be understood as cumulative business context of some compound message component, while the *Effective business context* should reveal message component's relevancy in a *Requested business context*.

The process of assigning contexts to a message profile and its components, and calculation of *Overall* and *Effective business contexts* is called *contextualization.*

3 E-UCM Approach – On-going Validation

Our initial validation [6] concluded that E-UCM methodology provides a powerful mechanism for business context presentation and expression. Our goal in this paper is to validate the algorithms that E-UCM proposed for both *contextualization* of existing message profiles and profile generation. Our approach to achieve that goal involves two steps. The first step is to assesses *Effective business context* as sufficient evidence of message component's relevancy or irrelevancy in the defined business context. The second step assesses the algorithms for message profile generation as capable of constructing a valid, message profile for a *Requested business context*. The key difference between those two goals is whether message profile structure is available or not. If it is available, profile refinement occurs for new integration scenarios. Otherwise, message profile structure should be generated using E-UCM algorithms on the basis of existing message profiles.

We made two important assumptions in completing the validation study. First, we assumed *contextualization* would be realized by a specialized service provider. Such a service provider would need domain knowledge needed for the development (or selection from an existing standard library) and use of message components. Such domain knowledge would typically be obtained through 1) business analysis of existing business processes and 2) information exchanges that rely on documentary standards and standard operating procedures.

Second, we assumed that there is one contextualized *Initial profile* upon which all others are based. To obtain the *Initial profile*, a service provider would construct a special kind of domain knowledge that contains an *initial collection of document components* and a *set of business rules* that govern the usage of those components. Using this domain knowledge, the service provider would generate an output in the form of a contextualized *Initial profile* with its components assigned intended business context, corresponding to the business rules.

Our validation process is shown in Figure 1. First three steps are prerequisite and the same for both validation goals, fourth and fifth steps reveal our validation goals, while the sixth step is an assessment of validation results. The third step gathers validation cases that will be used to accomplish our, two, identified, validation goals. Each validation case is described using *Requested business context*.

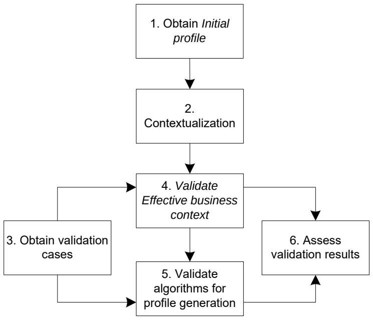

1. Obtain *Initial profile*
2. Contextualization
3. Obtain validation cases
4. *Validate Effective business context*
5. Validate algorithms for profile generation
6. Assess validation results

Figure 1 Validation process

3.1 Validation of *Effective business context*

The input for this validation step is semi-contextualized *Initial profile* with defined *Assigned business context* and calculated *Overall business* context and a set of validation cases. The output is a set of identified message profiles obtained through refinement of an *Initial profile*. The first step in our validation approach is to put the *Initial Profile* to use in a variety of validation cases, both boundary and normal, where business context is adequately and precisely expressed. For each validation case, *Effective business contexts* for each component will be calculated in order to identify irrelevant components. *Effective business context* is calculated as intersection between component's *Overall business context* and business context in which corresponding component is supposed to be used (*Requested business context*). Intersection is conducted for each business context category separately.

In general, there are three possible outcomes: *Effective business context* is equal to null, *Effective business context* is narrower than component's *Overall business context* and *Effective business context* is equal to component's *Overall business context*. *Effective business context* can be equal to null if intersection of the *Overall business context* and *Requested business context* expressions across any of their business context categories gives an empty set. This outcome informs us that the corresponding component or field is not relevant for the *Requested business context*. Other two outcomes inform us that the corresponding component is relevant, but it might undergo additional considerations (because *Effective business context* of the component is narrower).

The refinement of an *Initial profile* results in the identification of new message profiles for a specific set of integration scenarios, where only components that are relevant for the *Requested business contexts* are included. Message profiles' refinement results in recalculation of *Overall business contexts*. New message profiles will have *Assigned business context* equal to *Requested business context*. Figure 2 presents a reusable Contextualization and profile refinement subprocess that can be applied to any domain. Since the calculation of *Overall business context* depends on message component's type it is represented currently as a subprocess and details are neglected.

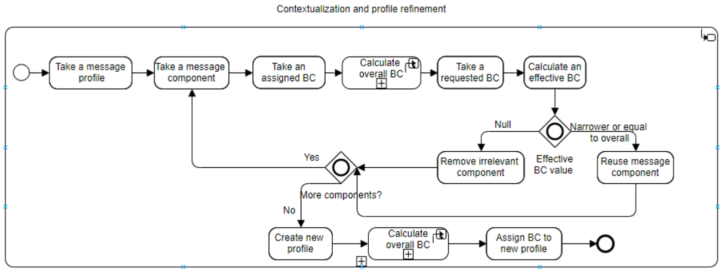


Figure 2 Contextualization and profile refinement process

3.2 Validation of algorithms for message profile generation

A set of refined message profiles, obtained from the previous step, and a set of validation cases provide the input for this validation step. An output is a set of generated new message profiles constructed using E-UCM proposed algorithms. In this validation step, we will take one by one validation case, and compare *Requested business context* with business contexts assigned to existing profiles that we got as the output from the previous validation step. If there is a match, we will reuse it, otherwise, we will employ E-UCM algorithms to obtain components, from the repository of contextualized ones, that are relevant for a new validation case. E-UCM proposes an algorithm with five processing steps. The first step gathers all components from an existing standard component library, while each of the next steps has the goal to filter message components applicable to the *Requested business context* regarding different matching indexes, such as complete or partial matching. Details of this algorithm are out of scope of this paper and will be discussed in our future work.

3.3 Results assessment

To achieve the identified goals, the structure of resulting profiles, both from the first and the second step, will be compared with expected outcomes by consulting human expert or developer. Important remark is that the fifth validation step is conducted under assumption that the output of the fourth step is valid.

4 Validation domain

We demonstrate our validation process using a case study from a Visa application domain. Afterwards, all conclusions will be verified in other domains, such as a Small-and-Medium-Enterprise (SME) and its procure-to-pay scenarios. We have identified three variables in Visa application domain: *Issuance country*, *Visa type* and *Applicant's country*. Moreover, we use these three variables as our business-context categories. Associated message profiles can include different combinations of these three. To simplify the complexity, we have fixed the *Issuance country=New Zealand* and *Visa type =Visitor*, respectively.

Since each country has its own, specific visa policies, we used a two-step process to construct a set of validation cases. First, for *New Zealand*, for *Visitor visa type* we identified a list of eligible countries whose citizens can apply for a visa. Second, for the same issuance country, for the same visa type we have identified a list of visa waiver countries. After these two steps are conducted, we can identify possible visa application submissions.

In the real life, New Zealand has one Visa Application Form (VAF) for Visitor visa type and this VAF, with the same structure, is offered to all applicants, no matter their country. According to our validation methodology setup, this *initial collection of document components* will be analyzed by the service provider, along with the set of New Zealand's specific business rules. As the output, we will get the contextualized *VAF Initial profile*. By studying the structure of visa applications for different visa types and different countries, we have concluded that there is a great chance that irrelevant fields will be proposed to a certain applicant. Korean application form has the field *Kanji Name* that is applicable for a few applicants' countries only. New Zealand's application form has the field *SWITCH card* which is the United Kingdom-specific credit card type, to name a few. These examples clarify the purpose of our first goal - to approve *Effective business context* as enough to identify relevant fields from the *VAF Initial profile*.

The output of this validation step would be a set of *VAF profiles* that would contain only components relevant to the *Requested business contexts*. Consequently, each applicant would get a customized VAF that contains only relevant set of fields, which will ultimately lead to less error-prone application process.

When a new visa application is recorded, first, we will compare *Requested business context* with business context assigned to existing profiles using the E-UCM business-context, matching algorithms. If there is a match, we will reuse it; otherwise, we will employ the E-UCM algorithm to construct a new message profile. That construction is based on relevant, obtained components from the repository of contextualized ones. The structure of resulting *VAF profiles* will be compared with expected outcomes by consulting human expert or developer. These conclusions contribute to the identified validation goals.

5 Discussion and Future Work

So far, first three validation steps have been conducted. Analysis of *Initial profile* and validation cases show that, in most situations, *Effective business context* can be used to identify irrelevant components from existing message profiles. However, there are some boundary scenarios where *Effective business context* is insufficient to make conclusions. We have identified the following types of unsuccessful validation cases. They justify the purpose of our validation process. From those example scenarios, E-UCM validation must be conducted prior to its industry adoption.

Time-dependent - Values assigned to business-context categories denote the current state of an object. We identified scenarios where it was necessary to remember the history, not just the current state. For example, keeping history of applicant's country transitions might be important to decide whether some field is applicable to the integration situation at hand. In this case, *Applicant's country* business-context category would be used to denote transition of the applicant's country, not just his current citizenship. Example transitions include changes in citizenship, holidays, temporary work, and student-exchange programs. In practice, this means that the result of *Effective business context* might vary through time. Hence, considering only one state, without considering history of change, may not be sufficient to communicate needed information for the relevant integration situation.

Unpredictable - There are scenarios that lack the logic needed to determine whether some field is relevant to the applicant or not. For example, field *Passport book number* can often be found in VAFs. The *Passport Book Number* may appear in a passport in addition to the *Passport Number*. Some countries may have this detail in some versions of their passports, while others may not. There is no traceable guideline that will inform us which passport versions contain this detail, so defining *Assigned business context* for such component would be a challenge. Further, it means that calculation of *Effective business context* might be impossible.

Insufficient - There are scenarios where business context does not give us enough information. For example, *Applicant's country* may not be enough because applicant's current residence or applicant's nationality, to name a few, may also affect appearance of some field. This means that the business-context categories that are chosen to describe business context are not enough; and, other relevant categories should be introduced. In practice, *Effective business context* can be calculated, but it can lead to conclusions that are inadequate for the specific integration situation.

If our validation process proves *Effective business context* as insufficient, which we believe is the case, future research will be needed. We plan to consider using new context categories and employing data mining techniques for business context definition to resolve those failing scenarios and enable adequate calculation of *Effective business context*. These approaches may help identify combinations of context categories expected to adequately define a specific business context [4]. Also, there is a possibility for extension of E-UCM methodology to capture history data through business context categories. A possible solution would be introducing state-machine diagrams for business context definition that would enable capturing multiple states (values for specific business context category) through time.

The described on-going validation covers algorithms that can be employed for profiles' generation when there is only one document type considered. Also, E-UCM offers other algorithms that can be applied for profiles' generation when there are two, or more, paired document types. Future research will consider validation of this aspect of E-UCM methodology as well.

Disclaimer

Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

References

1. The Open Applications Group Inc. https://oagi.org
2. The National Institute of Standards and Technology (NIST). https://www.nist.gov
3. Ivezic, N., Ljubicic, M., Jankovic, M., Kulvatunyou, B., Nieman, S., & Minakawa, G. (2017). Business process context for message standards. CEUR Workshop Proceedings, 1985, 100–111.
4. Jelisic E., Ivezic N., Kulvatunyou B., Anicic N., Marjanovic Z. (2019) A Business-Context-Based Approach for Message Standards Use - A Validation Study. In: Welzer T. et al. (eds) New Trends in Databases and Information Systems. ADBIS 2019. Communications in Computer and Information Science, vol 1064. Springer, Cham
5. Jelisic, E., Ivezic, N., Kulvatunyou, B., Jankovic, M., & Marjanovic, Z. (2019, September). A Two-Tiered Database Design Based on Core Components Methodology. In European Conference on Advances in Databases and Information Systems (pp. 350-361). Springer, Cham.
6. Elena Jelisic, Nenad Ivezic, Boonserm Kulvatunyou, Scott Nieman, Hakju Oh, Nenad Anicic, Zoran Marjanovic, Knowledge representation for hierarchical and interconnected business contexts, Enterprise Interoperability IX, Tarbes, France, 17. - 20. Nov, 2020
7. Novakovic, D. (n.d.). *Business Context Aware Core Components Modeling*. Retrieved September 28, 2018, from Publikationsdatenbank der Technischen Universität Wien: https://publik.tuwien.ac.at/
8. UN/CEFACT Core Components Technical Specification CCTS, version 3.0., https://www.unece.org. Last accessed 10 March 2020

High-Resolution Biochemical Activity Measurements with Commercial Transistors

Seulki Cho[1], Son T. Le[2,3], Curt A. Richter[2] and Arvind Balijepalli[1]

[1]Biophysics Group, Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

[2]Advanced Computing Group, Physical Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA

[3]Theiss Research, La Jolla, CA 92037, USA.

ABSTRACT

We demonstrate that single-gated, commercially-sourced, field-effect transistors (FETs) operated with a lock-in amplifier (LIA) under closed-loop control can achieve an average pH resolution of 9×10^{-4}. This performance represents an ≈8-fold improvement over previous FET measurements[1]. The pH sensitivity was found to be ≈56 mV, consistent with the Nernst potential at room temperature. The precision of our approach makes it ideally suited for sensitive bioanalytical measurements. We show that the technique can be applied to measure the therapeutic efficacy of small polypeptide molecules that regulate enzymes implicated in Alzheimer's disease.

KEYWORDS: Single-gate FET, pH measurement, PID controller, Lock-in amplifier, TiN, Sensitivity, Resolution

INTRODUCTION

The demand for high-performance biochemical measurements has dramatically increased to support applications such as biomanufacturing, marine ecology, and DNA sequencers[2]. Electronic measurements that leverage the field-effect are particularly attractive because they allow real-time signal detection without the need for specialized labeling of biomolecules. Their small size, low-cost, and scalability by leveraging foundry processes can allow a straightforward route to commercial deployment. FET-based biochemical measurements can allow the direct measurement of ions concentration (H^+ or OH^-) in a solution when paired with an appropriate sensor. As an example, measurements of pH can allow rapid quantification of enzymatic activity[3]. Improving the resolution of these measurements will allow the sensors to operate under physiological conditions. Therefore, many biochemical research groups are actively developing devices and methods to improve the sensing performance. Here, we demonstrate that commercial single-gate FETs combined with novel signal processing can allow nearly an order of magnitude improvement in pH resolution over conventional FET measurements.

EXPERIMENTAL

A single-gate commercially sourced n-channel FET was soldered onto a printed circuit board. The initial electrical characterization of the devices was performed with a semiconductor parameter analyzer. The pH measurements were performed in a remote configuration by connecting a TiN pH sensor connected to the top-gate metal contact using a shielded coaxial cable. The pH sensor was comprised of tungsten needle coated with 100 nm TiN thin film using sputter deposition[4]. The measurements were performed using two configurations. Figure 1a shows the schematic of single-gate FET measurement with a proportional-integral-derivative (PID) controller, operated in a constant current mode[1]. The PID controller was operated to maintain the drain current, I_D, at a constant value by continuously summing the controller voltage (V_{PID}) with the voltage from the pH sensor (V_{pH}). Figure 1e illustrates the schematic of single-gate FET measurement that combines a PID controller with phase sensitive detection using a lock-in amplifier (LIA). Similar to the measurement setup in Figure 1a, the FET operated in a constant current mode. The LIA allowed improved measurements of weak signals at a specific reference frequency and phase to improve the overall signal-to-noise ratio (SNR).

RESULTS AND DISCUSSION

Figure 1a describes single-gate FET measurement with PID control. The pH sensitivity (dV_{PID}/dpH) was determined to be ≈58.7 mV (R^2=0.99), consistent with the Nernst potential of 59 mV at room temperature from the time-series measurements shown in Figure 1b (*inset*). The time-series measurements also allowed the estimation of the pH resolution to be $\Delta pH = (7.2 \pm 0.3) \times 10^{-3}$ at a bandwidth of 10 Hz. This value was 3 times higher than when the single-gate FET operated in an open-loop without PID controller[1]. The improved pH resolution allowed high-resolution measurements of the effect of a polypeptide therapeutic, p5, on the regulation of the pathological enzyme complex, Cdk5/p25, shown in Figure 1c. Importantly, the measurements were made under physiological conditions with appropriate buffering conditions. As seen from Figure 1d, upon addition of the 24 amino acid polypeptide p5 we observed a significant inhibition in Cdk5/p25 activity as is evident from the change in V_{PID}. The observed p5-based inhibition of Cdk5/p25 activity had a threshold of [p5]=0.7 µM, consistent with literature values[5]. The performance of the measurements were improved by using a LIA in conjunction with PID control. In this configuration, the sensitivity of the FET was ≈56 mV (R^2=0.99) as seen from Figure 1f. The average resolution of FET extracted from the time-series measurements in Figure 1g improved 8-fold to an average of 9 x10^{-4} (Figure 1h). This improved resolution will allow enzymatic measurements at sub-physiological concentrations, allowing new tools for drug discovery and diagnostics of neurological disease.

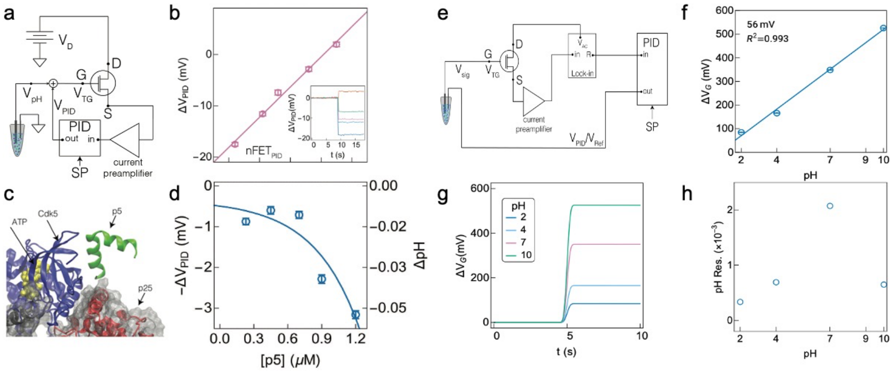


Figure 1: (a) Biochemical measurement using FET with PID control. (b) Response of ΔV_{PID} under different pH conditions. (inset) Time-series data as a function of pH. (c) p5 interactions with the Cdk5/p25 complex. (d) Response of ΔV_{PID} as a function of p5 concentration ([p5]). (e) Schematic of oprating a FET with a LIA and PID controller to improve measurement resolution. (f) Response of ΔV_{PID} uder different pH conditions. (g) Time-series under different pH condition relative to a reference potential. (h) pH resolution under a wide range of pH conditions.

CONCLUSION

We show that FET-based biochemical measurements can be substantially improved by combining commercially-sourced devices with PID and LIA techniques. The observed pH resolution was found to be ≈8-fold higher than conventional measurements where the FET was operated in an open-loop. The substantially improved resolution demonstrated by our results can lead to new tools for robust therapeutic development and diagnostic tools for use in neurological disease research.

ACKNOWLEDGEMENTS

We would like to thank Sufi Zafar from IBM for providing us with TiN needles. S. T. L. acknowledges support by the National Institute of Standards and Technology grant 70NANB16H170.

REFERENCES

1. S. T. Le, M. A. Morris, A. Cardone, N. B. Guros, J. B. Klauda, B. A. Sperling, C. A. Richter, H. C. Pant, and A. Balijepalli, *Analyst* 145, 2925 (2020).
2. S. Vigneshvar, C. C. Sudhakumari, B. Senthilkumaran, and H. Prakash, *Front. Bioeng. Biotechnol.* 4 (2016).
3. S. T. Le, N. B. Guros, R. C. Bruce, A. Cardone, N. D. Amin, S. Zhang, J. B. Klauda, H. C. Pant, C. A. Richter, and A. Balijepalli, *Nanoscale* 11, 15622 (2019).
4. S. Zafar, M. Khater, V. Jain, and T. Ning, *Appl. Phys. Lett.* 106, 106 (2015).
5. Y. L. Zheng, N. D. Amin, Y. F. Hu, P. Rudrabhatla, V. Shukla, J. Kanungo, S. Kesavapany, P. Grant, W. Albers, and H. C. Pant, *J. Biol. Chem.* 285, 34202 (2010).

CONTACT

* Arvind Balijepalli; phone: +1-301-975-3526; arvind.balijepalli@nist.gov

"It's the Company, the Government, You and I": User Perceptions of Responsibility for Smart Home Privacy and Security*

Julie Haney$^{\alpha}$, Yasemin Acar$^{\alpha\beta}$, and Susanne Furman$^{\alpha}$,
$^{\alpha}$*National Institute of Standards and Technology*,[†] $^{\beta}$*Leibniz University Hannover*
{julie.haney, susanne.furman}@nist.gov; acar@sec.uni-hannover.de

Abstract

Smart home technology may expose adopters to increased risk to network security, information privacy, and physical safety. However, users may lack understanding of the privacy and security implications. Additionally, manufacturers often fail to provide transparency and configuration options, and few government-provided guidelines have yet to be widely adopted. This results in little meaningful mitigation action to protect users' security and privacy. But how can this situation be improved and by whom? It is currently unclear where *perceived responsibility* for smart home privacy and security lies. To address this gap, we conducted an in-depth interview study of 40 smart home adopters to explore where they assign responsibility and how their perceptions of responsibility relate to their concerns and mitigations. Results reveal that participants' perceptions of responsibility reflect an interdependent relationship between consumers, manufacturers, and third parties such as the government. However, perceived breakdowns and gaps in the relationship result in users being concerned about their security and privacy. Based on our results, we suggest ways in which these actors can address gaps and better support each other.

1 Introduction

While early adopters of IoT smart home technology have typically been more technically savvy, smart home devices are increasingly being purchased by non-technical users [31] who may not understand the technology's privacy and security implications. Within the current dynamic threat and technology environment, the uptick of smart home technology adoption may expose users to increased risks to their network security, privacy of their information, and quite possibly their physical safety [26]. In addition, global surveys have identified that security and privacy are significant concerns among both IoT adopters and non-adopters [9, 49], and that consumers would like more information about security and privacy when purchasing devices [33]. Therefore, it is imperative that smart home consumers be empowered to protect the security and privacy of their devices while still being able to enjoy the benefits of the technology. This would result in consumers feeling more comfortable with their devices and encourage additional adoption among those who currently have concerns.

Unfortunately, smart home devices may fail to provide transparency of privacy and security protections and may lack adequate security and privacy controls [24], while manufacturers may be unsure as how best to implement these [25]. Generally, third-party guidance on desirable privacy and security controls has not yet entirely converged and is not currently widely adopted since many of these efforts are nascent and reflect in-progress work.[1] In combination with users' lack of in-depth understanding of smart home device technology, privacy, and security, the result is limited meaningful mitigation actions being taken to protect consumer security and privacy [1, 32, 42, 49, 66]. For example, some users leave the room to have sensitive conversations out of earshot of the technology, unplug devices, or tape over cameras.

In order to create meaningful and effective privacy and security controls, interfaces, guidelines, and other resources to support users, it is important to understand who users believe are the responsible parties for privacy and security. Responsibility can be viewed as being active: "the state or fact of having a duty to deal with something."[2] A better understanding of perceptions of responsibility and framing within the context of duty/obligation might shed further light on what actions users are willing and able to take on their own versus

[*] Authors' extended version of paper to appear in 2021 USENIX Security Symposium

[†] Certain commercial companies/products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the companies/products identified are necessarily the best available for the purpose.

[1] E.g., NISTIR 8259 was published in May 2020 [23]; ENISA published the updated *Good Practices for Security of IoT* [22] in November 2019; the UK published *Code of Practice for Consumer IoT Security* [17] in October 2018.

[2] https://www.lexico.com/en/definition/responsibility

Haney, Julie; Acar, Yasemin; Furman, Susanne M. "It's the Company, the Government, You and I: User Perceptions of Responsibility for Smart Home Privacy and Security." Paper presented at USENIX Security 2021, Vancouver, CA. August 11, 2021 - August 13, 2021.

which functions they feel are the duty of or would be better suited to others. Knowing the will of the consumer may then put more pressure on others to take action. We also consider that responsibility may be perceived in a more negative light as "the state or fact of being accountable or to blame for something."[3] Viewing responsibility through this lens may reveal areas of discomfort. These areas of discomfort could illuminate gaps that need to be filled in order to provide a more private and secure smart home experience and make adoption more palatable. However, it is currently unclear where users think responsibility for smart home privacy and security lie.

To address this gap, we uncovered perceptions of responsibility during a semi-structured interview study of 40 smart home users by seeking to answer two research questions:

RQ1: Who do users believe is responsible for the privacy and security of their smart home devices?

RQ2: What is the relationship, if any, between perceptions of responsibility, concern, and taking mitigative action?

Our study revealed that user concerns about the possibility of undesirable security and privacy situations (e.g., as found in [56,66]) can stem from the perception of insufficient controls on manufacturers and inadequate user support. We found that users primarily assign privacy and security responsibility to three actors or a combination of those - smart home owners (personal responsibility), manufacturers, and government/regulatory bodies - with manufacturers being most frequently held responsible. Responsibility is often viewed as being an interdependent relationship between those actors in the pursuit of robust smart home privacy and security. Part of this relationship relies on actors taking voluntary action (e.g., users configuring security options) and supporting the others in their goals (e.g., a manufacturer providing security tips to consumers). However, when a user is either unwilling or unable to take necessary action, participants desired better information and built-in protection by manufacturers, facilitated by the government. When manufacturers do not use privacy and security standards or support privacy/security controls, standards or guidance can help them target a privacy/security baseline, with "checks and balances" (e.g., regulations, certification) enforcing action.

Our study makes several contributions:

- We provide novel insight into where smart home users place responsibility for the privacy and security of their devices and how those perceptions may relate to concerns and implementation of mitigations. We identify a theme of an interdependent relationship between users, manufacturers, and the government/third parties.
- Our findings extend prior literature related to perceptions of privacy/security responsibility for conventional technology into the smart home domain.
- We give practical guidance for how users, manufacturers, and government/third party organizations might support each other by filling current gaps.
- We suggest future research directions to address how best to enhance the interdependent relationship necessary for smart home privacy and security.

2 Background

To help frame our smart home privacy and security study, we describe prior research and background information related to privacy/security perceptions, smart home privacy/security, responsibility, and third-party efforts.

2.1 Related Work

2.1.1 Privacy and Security Perceptions

Prior research on privacy perceptions can serve as a foundation when exploring user beliefs and opinions of smart home privacy. Researchers have suggested the existence of a "privacy paradox" [2,7] in which, although users often state that they care about privacy, they may fail to mitigate privacy risks and choose to use privacy-violating technology. Users may also willingly or reluctantly trade privacy and security for convenience and perceived benefits [2,47,48]. One study suggests that users value privacy more when they have it than when they do not, i.e., efforts to re-establish privacy may be less spirited than staying private in the first place [3]. Default settings and hard-to-navigate configuration options also contribute to behavior that does not preserve privacy [46]. Furthermore, privacy policies are often mistakenly assumed to contain the promise to respect user privacy or understood as implicit recommendations [41]. The concept of "privacy resignation" in response to repeated privacy violations has also been identified [52].

We also turn to prior literature on perceptions and security mitigations employed with traditional information technology (IT) and online applications as a potential basis of comparison. Typical, non-technical end users rarely view security as a primary goal when interacting with technology, often lack security knowledge, and have low self-efficacy when it comes to taking security-related action [54]. This is opposed to security experts who have very different ideas of which actions help with online security [39,58]. Wash and Rader [62] surveyed U.S. internet users and found that those with weakly held beliefs about viruses and hackers were the least likely to take protective actions. Stanton et al. [54] discussed "security fatigue," a weariness towards security when it becomes too burdensome. Herley [36] similarly claimed that users may ignore security advice due to being overwhelmed by the sheer volume of advice, viewing security as being a high cost to themselves, and because they perceive security actions to be

[3] Ibid.

inadequate in the face of myriad threats. West et al. [63] examined why people make poor security decisions, finding that the tendency to satisfice, cognitive biases, time pressures, and inattentional obliviousness contribute to this.

In this paper, we explore whether users' general views of privacy and security found in the literature are reflected in the perceptions of privacy/security responsibility for a specific technology (smart homes).

2.1.2 Smart Home Security and Privacy

In recent years, many researchers have examined smart home privacy and security from a user perspective. In this section, we highlight several relevant efforts that identified user perceptions and experiences that can be confirmed or extended in our own study. Early work pointed out a lack of transparent privacy controls in smart home devices [61]. A subsequent study identified additional challenges and tensions in smart home hubs, including security and privacy issues [44].

Research and industry surveys have shown that security and privacy concerns can be barriers to adoption of smart home devices [11, 21, 57, 64]. For example, Lau et al. [42] found that some non-users are privacy conscious and distrustful of privacy and security of smart home devices and their manufacturers, and that smart home devices generally cross these non-users' privacy thresholds.

Even adopters have privacy and security concerns. For example, Sanguinetti et al. [51] found that owners of smart home devices were just as concerned as those who chose not to purchase the devices. Malkin et al. [43] observed that users express concern about smart home speaker recordings and reject the use or sharing of recordings for purposes other than voice commands because of a violation of contextual integrity (i.e., not adhering to user expectations of how data flows and is used for a specific service). Users also have complex, but incomplete threat models, which include a general sense of being surveilled by manufacturers or the government and the possibility of being attacked by hackers, while lacking awareness of botnets and the sale of inferred data [1, 21, 67]. Users were generally more concerned when the privacy of children was at stake [4, 43].

Smart home users also express that they lack information to evaluate device privacy and security features. Emami-Naeini et al. [21] found that, although participants ranked privacy and security as important factors when purchasing IoT devices, information was difficult to find. This was also confirmed by researchers at the U.S. National Institute of Standards and Technology (NIST) who found that open-source security information for smart home devices often lacked specificity or was unavailable [24].

Multiple studies found a lack of substantive mitigation actions to address security and privacy concerns for various reasons, including lack of agency, lack of option availability, and trust in other entities to take action [1, 32, 42, 56, 66]. Adopters may also fail to take action because they typically have higher tolerances for privacy violations, willingly or reluctantly accept the trade-off in exchange for the convenience and utility offered by smart home devices, and often express that they have "nothing to hide" [42, 56].

Other researchers identified privacy and security options desired by users. In a co-design exercise, Yao et al. [65] found that data localization and a private mode were among desired items for privacy protections. Haney et al. [32] identified wishlists for both privacy and security mitigations, which included more transparency about data collection and use and easy-to-configure options. However, availability of options must be balanced with usability, as expressed by Colnago et al. [12] who found that, while participants desire more control over their data and privacy settings, they are concerned about being overloaded with configuration options and "notification overload."

Several studies investigated the use of smart home devices in multi-user homes, finding power imbalances in that secondary users often have less agency in purchase and configuration and use decisions, which creates a potential for abuse [28, 42, 66]. These findings are corroborated by Huang et al. [37], who observed that users of multi-user devices adopt all-or-nothing mitigation strategies similar to mitigations against external actors, and desire more control options over their data. Tabassum et al. [57] found that users desire sharing options with people outside their home to increase their security. Based on a 2018 online study, He et al. suggested that smart homes need granular configuration options based less on device type and more on user type (e.g., neighbor vs. spouse) [35]. On the manufacturer side, Chalhoub et al. interviewed smart camera designers, and found that user experience (UX) is considered important in communicating privacy configurations, but is under-utilized when it comes to security [10]. While prior studies identified smart home privacy and security concerns and mitigations, to the best of our knowledge, none explored *perceptions of responsibility* in detail. This is a gap our research hopes to address.

2.1.3 Perceptions of Responsibility

As a possible comparison point to our findings related to responsibility of *smart home* security and privacy, we look to prior work addressing general security and privacy responsibility. Past research has shown that consumers often feel that security is the responsibility of a third party (for instance, the government, vendors, or IT professionals) and may delegate security decisions because they feel they lack knowledge and technical skills to take action [27, 30]. From a privacy perspective, Renaud at al. [50] explored why end-to-end email encryption solutions have not been widely adopted. They found that, although participants were privacy aware, they were often not overly concerned enough to take additional action, partially because they abdicated responsibility to service providers that they felt were better equipped. Bandyopadhyay [5] proposed a theoretical framework to explore factors influencing privacy and security concerns of consumers who use the internet. He suggested that there is a consumer trust problem which necessitates increased assurance that security and privacy are being protected. Therefore, the responsibility of assurance was viewed as three-fold, falling on governments, vendors, and, to a lesser degree, consumers. Dogruel and Joeckel [19] interviewed U.S. and German smartphone users and found that most felt the responsibility for privacy protection lies primarily in their own hands. While some participants assigned third party responsibility to government and commercial entities, most believed both carry at least some responsibility for privacy. German participants were much more likely to desire government intervention in the case of privacy, for example by setting minimum privacy standards and establishing legal frameworks. U.S. participants, however, were more likely to place accountability with commercial entities.

A global Mozilla survey of close to 190,000 people asked "Who is most responsible for protecting the online safety, privacy, and security of the connected apps and devices you own?" [9]. Thirty-four percent of respondents placed responsibility on the makers of apps and devices, with roughly the same percentage saying that it was up to them. Twenty percent selected government. The survey also revealed variances in responsibility perceptions among different countries. For example, respondents from Mexico and the U.S. were much more likely to claim personal responsibility (41% and 43%) and less likely to put most responsibility on the government (13% and 12%) as compared to those from other countries.

While these prior studies examined perceptions of responsibility, none focused on smart home devices. It is unclear as to whether responsibility for smart home devices is viewed differently than traditional online or information technology, potentially because of inherently unique characteristics of the devices, such as them being always on and collecting data within highly personal and private spaces. Our study begins to address this unknown.

2.2 Third-Party Efforts

Government, regulatory bodies, non-profits, and other certification authorities have demonstrated initiative in protecting consumers' digital privacy and security, with differing levels of success. Recent developments in privacy-protecting laws reflect that some responsibility for keeping user data private is being shifted from users to corporations via government intervention. For example, the European Union (EU) enacted the General Data Protection Regulation (GDPR) [60], which provides individuals with rights related to the collection and storage of their personal data and requires that developers implement privacy by design. In the U.S., the state of California recently implemented the California Consumer Privacy Act (CCPA) [55], a statute that addresses online privacy and states that a consumer has rights regarding transparency of data collection and the right to request that their data not be sold and be deleted. Reactions and implementations for these regulations have been mixed since privacy may be viewed as a conflict between allowing the free market to trade data as a commodity and empowering end users to control their own data. With respect to GDPR, while some vendors have added configuration options, many are still difficult to navigate for average users. Other vendors block access to their services when accessed from within the EU to avoid having to comply [16].

With respect to IoT, several industry, government, and non-profit organizations have issued voluntary security guidance for manufacturers, most of which is too new to have been widely adopted. Recent government guidance includes NIST's *Foundational Cybersecurity Activities for IoT Device Manufacturers* [23] in the U.S., the European Union Agency for Cybersecurity (ENISA)'s *Good Practices for Security of IoT - Secure Software Development Lifecycle* [22], and the United Kingdom (U.K.)'s *Code of Practice for Security of IoT* [17]. Industry consensus groups have also provided privacy and security baseline resources for manufacturers, for example, the Internet of Things Privacy Forum [38], IoT Security Foundation [40], and the Council to Secure the Digital Economy [14]

Recently, there has also been considerable attention and advocacy for IoT product security and privacy labels as both an aid to consumers and way to increase manufacturer transparency and accountability [18, 33, 53]. For example, the Underwriters Laboratory (UL) now provides an IoT security rating backed by a standardized process to evaluate security aspects of smart products [59] and the wireless industry association implemented the CTIA IoT Cybersecurity Certification Program [15]. Carnegie Mellon University proposed IoT security and privacy labels based on studies of consumers and experts that suggested that labels could aid in consumer purchase decisions while holding manufacturers accountable for product privacy and security implementations [20, 21].

3 Methods

Between February and June of 2019, we conducted an exploratory, semi-structured interview study of 40 smart home users to understand their perceptions of and experiences with the devices. This paper describes a subset of collected data which is novel to prior smart home research and centered on user perceptions of privacy and security *responsibility*. The study was approved by our institution's research protections office. Prior to data collection, participants were informed of the study purpose and how their data would be protected. Data were recorded without personal identifiers (using generic identifiers such as P10_A) and not linked back to individuals.

3.1 Participant Recruitment & Demographics

To be eligible for the study, participants had to be adult users of smart home devices. We hired a consumer research company to recruit general public participants, who were compensated with a $75 prepaid card. Prospective participants were members of the consumer research company's research panel, a database comprised of over 6,000 participants located in the Washington, D.C. metropolitan area in the U.S. who had agreed to be contacted about consumer research opportunities. The recruitment company emailed a subset of 444 members of the research panel, selected for demographic diversity. They also recruited via social media posts and requested direct referrals.

To determine eligibility, those interested in the study first completed an online screening survey about their smart home devices, their role with the devices (e.g., administrator, user), professional background, basic demographic information (age, gender), and number of household members. After reviewing the screening information, we purposefully selected participants for interviews if they had two or more different smart home devices for which they were an active user (as opposed to being a bystander). We did this to engage with users who actually had *smart homes*, which we define as using multiple, diverse smart home devices, as opposed to those with only one individual smart home device. Smart TVs were not included in this initial count (but were addressed in the interviews) because most TVs now come with smart functionality and do not necessarily represent a deliberate choice to purchase a smart device.

We ultimately selected and interviewed 41 individuals. Despite a review of the screening questionnaire, during the interview, one participant (P5) was found not to have any smart home devices, so was removed from the study.

We defined smart home devices as being networked devices in the following categories, which were developed after consultation with IoT experts in our institution and used in the screening survey to focus responses. Number of participants with each type of device is indicated in parentheses.

- **Smart security (n=35):** e.g., security cameras, motion detectors, door locks
- **Smart entertainment (n=38):** e.g., smart televisions, speakers, streaming devices, connected media systems
- **Home environment (n=38):** e.g., smart plugs, energy consumption monitors, lighting, thermostats, smoke and air quality sensors
- **Smart appliances (n=15):** e.g., smart refrigerators, coffee pots, ovens, washing machines
- **Virtual assistants (n=36):** e.g., voice-controlled devices such as Amazon Echo/Alexa and Google Home

Initially, although not a major focus of this project, we also wanted to examine potential differences between smart home users living in the same household. Therefore, the survey was administered over the phone to another household member if interested. This recruitment only yielded four additional participants, so we ultimately decided not to pursue this vein of comparison. Since few participants were recruited in this way, it is unlikely that their opinions caused undue data bias, especially since most had different perspectives from their housemates.

Of the 40 participants, 32 had installed and administered the devices (indicated with an A after the participant ID), and eight were non-administrative users of the devices (indicated with a U). Twenty-two (55%) were male and 18 (45%) were female. The majority (70%) were between the ages of 30 and 49. Participants were highly educated with 18 (45%) having a master's degree or above and another 20 (50%) with a bachelor's degree. Thirty-four participants lived in multi-person households, with four couples among the participants (interviewed individually). All but one participant had three or more individual smart home devices, with 34 having devices in three or more categories. Refer to Appendix A for detailed participant demographics.

3.2 Data Collection

In addition to the screening survey responses, our data consisted of transcripts from 40 in-person, semi-structured interviews lasting on average 41 minutes. All interviews were audio recorded and then transcribed by a third party service provider. We chose semi-structured interviews over other methods, such as surveys, due to the exploratory nature of our investigation. Interviews afforded a greater richness of data, the ability to ask follow-up questions to more deeply explore participant responses, and the opportunity for participants to add other relevant information not explicitly targeted [13].

To develop our interview protocol, we conducted an extensive review of prior literature and market research up through 2018 to understand recent research, trends, and the state-of-the-art in smart home technologies. We also examined existing smart home devices ourselves to understand their usage. Based on these investigations, we crafted questions to address research gaps and explore multiple aspects of smart home device ownership and usage, including privacy and security. We asked an IoT domain expert to review our interview questions to ensure we were using correct terminology and considering appropriate facets of smart home ownership and use. We then piloted the interview protocol with four smart home owners from our institution (two device administrators and two non-administrators/users) to determine the face validity of questions and language. Pilot participants were not compensated. We made minor adjustments to the interview instrument based on feedback from the content expert and the pilot experience. Because modifications were only minor to improve clarity and comprehension, the pilot interviews were included in the final data set.

Interview questions addressed several areas in the following order: understanding of smart home terminology; purchase decision process; general use; general concerns, likes, and dislikes; installation and maintenance; privacy; security; and safety. Interview questions can be found in Appendix B. During the interviews, we differentiated between privacy and security by giving the participants definitions and examples of what each term meant. Security concerns relate to safeguarding of data/devices while privacy is safeguarding user identity (which can be gleaned from certain types of data). In this paper, we focus only on collected data pertaining to privacy and security *responsibility* since this topic has not yet been explored in detail by other researchers. Note that participants may have mentioned privacy and security responsibility concepts throughout the interview (for example, when asked if they had any hesitations prior to device purchase), not just during the designated privacy and security sections.

We interviewed until we reached two conditions. First, we monitored for theoretical saturation, the point at which no new ideas emerge from the data [13]. We also wanted to ensure we had a participant sample with a diverse set of smart home devices to account for potentially different experiences depending on the types of devices.

3.3 Data Analysis

Data analysis included both deductive and inductive coding practices, which allowed for an emergence of core concepts. Analysis of the interview transcripts began with the development of an *a priori* code list based on the research questions. Using the initial code list, each of the three research team members individually coded a subset of four interviews (4936 lines, 214 minutes of audio), then met as a group to discuss code application and develop a codebook. The final codebook addressed all data concepts (e.g., purchase, installation, usability, privacy, security, safety). All codes were "operationalized," which involves formally defining each code to ensure understanding among all coders. The codebook for privacy and security concepts informing this paper are included in Appendix C.

Using the codebook, we then coded the remaining interviews independently, with each transcript coded by two researchers and one primary coder (the first author) coding all interviews. Each pair of coders then examined and resolved differences in code application. In accordance with the recommendation of qualitative methodologists (e.g., [6, 45]), we focused not just on agreement but also on how and why disagreements in coding arose and the insights afforded by subsequent discussions. This focus was especially valuable in pursuing alternate interpretations of the data given the diverse perspectives of our multidisciplinary research team. When disagreement occurred, we discussed as a group to reach consensus. In rare cases where agreement could not be reached, the primary coder made the final decision.

Throughout the data analysis phase, we progressed to the recognition of relationships among the codes and examined patterns and categories. We met regularly as group to discuss our interpretations and emergent ideas. This process allowed for the development of central concepts, including the topic of this paper: perceptions of privacy and security responsibility as an interdependent relationship.

3.4 Limitations

As with any interview study, participant responses are subject to recall, self-report, and social desirability biases. In addition, our study only captures perceptions of smart home adopters of multiple devices, so does not adequately capture those of limited adopters or non-adopters. The participants, who were generally highly educated professionals in a high-income metropolitan area, may not be fully representative of the smart home user population in the U.S. However, our sample appears to mirror smart home adopters characterized in prior industry surveys [29]. We also acknowledge that U.S. smart home users may have different privacy and security attitudes from those in other countries, for example, due to political or cultural factors related to privacy expectations and tolerance. However, since other regions in the world, such as Europe, lag behind North America in terms of smart home market penetration and maturity [8], our findings may identify potential areas that other countries may want to consider as adoption increases. These limitations could be addressed with replication of this study in other countries or a global quantitative survey informed by the results of our study.

Since the smaller sample common to qualitative research does not lend itself to generalizability, we did not perform analysis to identify differences based on demographics (e.g., gender, age). We also did not differentiate responsibility based on device type but rather asked about general perceptions. We plan to explore the effect of demographic characteristics as well as per-device differences in a follow-up quantitative survey administered to a larger sample.

4 Results

In this section, we report results about perceived responsibility for smart home privacy and security. Example quotes from participants are provided throughout. Counts are provided in some cases, not as an attempt to distill our qualitative data to quantitative measures, but rather to illustrate weight or unique cases.

We first provide a brief overview of the privacy and security concerns and mitigations voiced by participants during the interviews. Although these concerns and mitigation strategies are not novel as compared to those identified in several of the studies cited in Section 2.1.2, we summarize our own

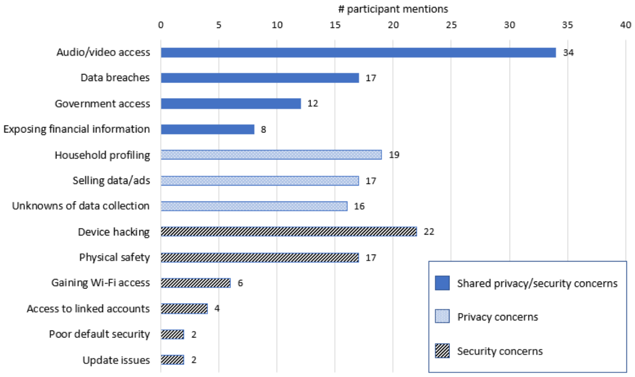


Figure 1: Participant concerns.

findings here in order to contextualize the focus of the paper: the assignment of *responsibility* for security and privacy.

4.1 Concerns and Mitigations

Early in the interview, we asked participants a general question, “What concerns, if any, do you have about the devices?” We later asked, “What are your concerns, if any, about how information is collected, stored, and used and who can see that information?" and “What are your concerns, if any, about the security of your devices?” In some cases, participants were personally concerned about privacy or security (28 for privacy and 26 for security) but to varying degrees. Several participants mentioned concerns that were expressed by others (e.g., family members, friends, media) but not personally held (4 for privacy, 6 for security). The most frequently mentioned concerns for both privacy and security in our study are summarized in Figure 1.

We also found evidence of lack of concern. In 24 cases, participants did not value the information collected by smart home devices, believing they would not be a worthwhile target. Therefore, they felt that there was a low probability that their devices would be hacked (5 participants). In addition, unconcerned participants often demonstrated privacy resignation [42] in which users believe that their data is already publicly available via other means and that there is nothing they can do about it (8 participants).

Privacy and security mitigations enumerated by participants were often simplistic or non-technical. Examples of simplistic mitigations include: setting a device app password, password-protecting the Wi-Fi network, and disabling the option to order items via virtual assistants. Non-technical mitigations included: not having sensitive conversations near virtual assistants, not placing devices with cameras or microphones in private rooms of the house (like bedrooms), or unplugging the device when not in use. Figure 2 shows the most frequently-

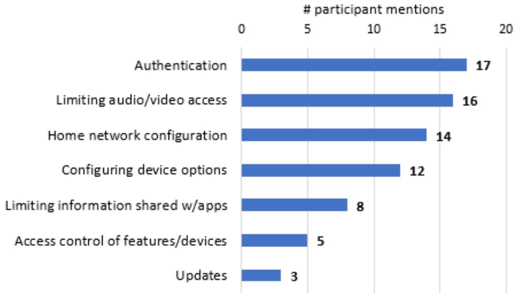


Figure 2: Shared privacy and security mitigations.

mentioned mitigations. Note that all of these were discussed at least once within both the privacy and security contexts.

We observed that being concerned about smart home privacy and security did not always translate into action. This inaction was due to several reasons. First, smart home device ownership was often viewed as a conscious choice to accept risks in exchange for perceived benefits, described as *“willful ignorance”* by P1_A. This same participant commented, *“It’s a trade-off. . . I know that it’s collecting personal data,. . . and I know there’s the potential of a security leak, but yet, I like having the convenience of having those things” (P1_A)*. Second, users may not be aware of available options or were not given options by the manufacturer. For example, one smart home user commented, *“I’ve been given very little methods to alleviate the concerns. Usually the description of the controls aren’t specific enough for me to alleviate my concerns” (P13_A)*. In addition, some do not have enough knowledge to be able to select and implement mitigations, especially security ones (8 participants). A participant said, *“I know it is password protected. That’s as far as my knowledge. I don’t know more than that. I’m not certified with cybersecurity” (P41_U)*. As with concerns, we also observed the influence of resignation as well as loss of control and fatalism, which are characteristics of security fatigue. One participant exhibited this resignation when he said, *“I just kind of assume if it exists, there’s a way to hack into it” (P18_A)*.

4.2 Responsibility

Participants were asked “Who do you think is responsible for protecting the privacy of information collected by your smart home devices?” and, later in the interview, “Who do you think is responsible for the security of your devices?” Participants may have also discussed concepts related to responsibility in response to other questions, e.g., those pertaining to concerns and “What kind of things would you like to be able to do with your devices, but haven’t, don’t know how, or are not sure that you can?” (see Appendix B for contributing questions in bold).

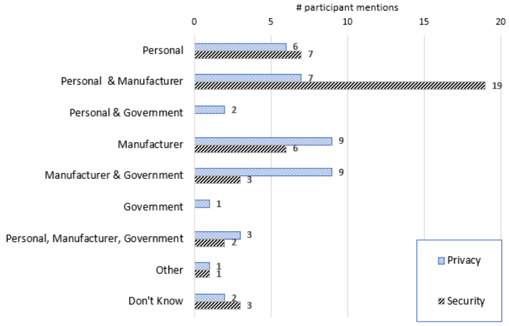


Figure 3: Perceptions of responsibility for smart home privacy and security.

Most responses fell into one of three categories or a combination of those: personal responsibility (smart home owners), device manufacturers, and government/regulatory bodies (see Figure 3). Two participants did not have an answer for privacy, and three did not have a response for security. One owner of a smart thermostat thought the power company was responsible for privacy, and one participant said internet service providers were partially responsible for security.

4.2.1 Personal Responsibility

Eighteen participants claimed at least partial personal responsibility for **privacy** (6 of those with sole responsibility). For example, P1_A expressed, *"It starts with us. We're bringing this device into our home."* Twenty-eight participants claimed some personal responsibility for **security** (7 with sole responsibility): *"It's on you to either put extra restrictions in place or just be okay with the fact that [a breach] is going to happen" (P8_A)*. Note that several participants placed responsibility on a housemate or spouse who was more involved with the devices. However, we considered personal responsibility as being that of smart home owners in general.

Eleven participants viewed personal responsibility as holding themselves accountable for accepting risks. For instance, personal privacy responsibility was often described as being implicit with device purchase and continued use. When asked who was responsible for privacy, a participant said:

> *"The owners. In my opinion, if you don't want stuff exposed, you shouldn't have those devices in your house to begin with. You're accepting a risk by taking those on in your home" (P35_A).*

Another commented, *"You buy the device and realize what you're getting yourself into… Buyer beware. Operate at your own risk" (P26_A).*

We also observed that viewing responsibility as personal could also be a justification for inaction in taking mitigation actions, even if privacy and security were concerns. In these cases, participants accepted personal blame for their own perceived deficiencies, such as not looking into what options were available, having incomplete threat models, or not taking the time to learn more about how to secure their devices or home networks. For example, P14_U believed device owners are to blame if they do not adequately secure their devices: *"I think that's probably a shared thing… A lot of people don't put secure passwords and stuff on their systems… People don't use the tools that are out there, like VPNs,… I think that's all responsibility of you."* Although P8_A believed he is solely accountable for the security of his smart home devices, he did not take many substantive mitigation actions because *"I'm not going to educate myself on network security… This stuff is not my forte. I'm very accepting to the fact that it is what it is."*

Conversely, participants who approached personal responsibility as an active, obligatory role were those who implemented mitigations above and beyond setting a password at installation and incorporated security and privacy considerations into their purchase decision-making process. Regarding the obligation to configure privacy settings, a smart home owner remarked, *"I feel like the default is always full access, so you have to really look for and pursue stricter settings" (P18_A)*. Especially in the case of security, responsibility was viewed as requiring some effort on behalf of users. For example, P15_A addressed most of his concerns by doing extensive research on the devices prior to purchase. He then only selected those he felt adequately implemented security and privacy protections, including *"good authentication, encryption, secure protocols being used."*

Some participants did not mention taking personal responsibility for smart home privacy and security (22 for privacy, 12 for security). We note that most of these participants did not explicitly deny responsibility, but rather assigned responsibility to other actors when asked. An older smart home user was one of the few to overtly abdicate responsibility when she said, *"I'll leave that to the next generation" (P38_U).*

The study results also revealed a disconnect between being concerned and accepting responsibility. Among those participants who accepted personal responsibility, the majority did express personal concern (13 concerned vs. 5 unconcerned for privacy and 20 concerned vs. 9 unconcerned for security). However, privacy concern did not necessarily mean that participants accepted responsibility (15 concerned did not accept responsibility for privacy vs. 13 that did). Being concerned with security was more likely to be associated with personal responsibility (20 accepting responsibility and 6 who did not).

4.2.2 Manufacturer Responsibility

As the most frequent response, 28 participants believed manufacturers share some responsibility for **privacy**, with nine of those assigning sole responsibility to manufacturers. For example, a participant remarked, *"Any single person who was involved in the creation of the product is responsible for what it does, including collecting information" (P30_U)*. Another felt that manufacturers *"have a responsibility to make sure that information is where it's getting sent to, who's getting it, and that it's safe, and it's not going to get taken away or stolen" (P32_A)*.

Thirty participants said manufacturers have at least some responsibility for **security** (only 6 for solely responsible). For instance, one participant who thought manufacturers are solely responsible said, *"I would say the manufacturer. I don't think they can expect all of us to be cybersecurity experts. That's why we bought the product" (P29_A)*. Another commented,

> *"[Manufacturers] are the prime people who are responsible for things they're making because we're not putting all the time, and energy, and money on building that stuff. So, we really don't know what is inside of this" (P9_A).*

The data revealed an attitude that manufacturers have an obligation to the buyers of their products to adequately protect their privacy and security, with this being part of an unstated manufacturer-consumer contract put in place at time of device purchase. One participant remarked, *"They need to do everything [since they are] taking so much money for all that" (P9_A)*. Another commented, *"If I'm going to buy your product, I think you owe it to me to not abuse that. I did give you money for it" (P29_A)*.

However, there were differing levels of confidence in whether manufacturers could adequately uphold this obligation. Participants who put their trust in manufacturers to protect their privacy and security often did so based on a perceived competence due to company size or reputation. For example, a user trusted larger companies to build secure devices: *"Maybe that's why I'm feeling a little more secure than not because I'm like, oh, this is a big company. If something happens, hopefully, they have the money to figure it out" (P6_U)*. One participant felt that it was beneficial for manufacturers to implement strong privacy and security measures because *"If they have a bunch of massive security breaches, people are going to stop buying their products. So our interests are aligned there" (P17_A)*.

Even though they placed responsibility on manufacturers, others expressed varying levels of distrust. Only 11 participants relied on manufacturer-supplied information when researching potential products, while 34 looked at other, often subjective online sources, such as customer reviews. While 10 participants believed data was sent to manufacturers for beneficial reasons (e.g., product improvement and tailoring to consumer habits), others felt that they were at the mercy of manufacturers who do not have consumers' best interests in mind, for example, believing manufacturers were purposely vague in terms and conditions statements so that consumer data could be more easily monetized. When asked if he ever reads any of the privacy agreements, P10_A said, *"I don't have much trust in what companies say they collect and don't collect. I think they collect what they can and use it."* Others felt that manufacturers were powerless to prevent data breaches and device compromise when up against a determined adversary. For example, a participant commented, *"I would say that I think they try to do a good job of being secure, but we see hacks all the time... I think that sooner or later they will get hacked" (P26_A)*.

In all of these cases, participants felt that manufacturers *should* have a duty to implement adequate security and privacy mechanisms but were not certain they *would* or *could*. However, manufacturers were still not exempt from being accountable or blamed if something should go wrong.

4.2.3 Government Responsibility

Fifteen participants thought that the government or some regulatory body was at least partially responsible for smart home **privacy**, with only one viewing government as being solely responsible. In general, participants viewed the government as having an obligation to protect its people from harm from security and privacy breaches. For example, a participant saw government regulation of smart home privacy as being associated with consumer safety:

> *"I think the other half of the responsibility goes on the government to protect your citizens... There's other safety precautions put in other industries. I don't see why that shouldn't be something applied to this industry as well" (P29_A).*

P31_A did not think the government would do the best job, but felt regulation had some benefit:

> *"We've got to do something to protect people's information, or at least make them more aware of what exactly is being utilized and sold, and having opportunities to opt-out, taking at least some steps."*

The assignment of government privacy responsibility was at times ironic because several participants also expressed that they believed the government was performing surveillance of citizens via smart home devices. Potential surveillance bothered some, but others were not concerned because they felt they were not doing anything illegal or of interest to the government. Even though P26_A thought the government was partially responsible for privacy, he remarked:

> *"I'd like to regulate our government, but that's not gonna happen. Right? I don't mean to sound so flippant, but I wish they would stop watching and collecting data, but that's not going to happen. It is what it is."*

Interestingly, while over a third of participants allocated at least partial responsibility for privacy on the government, there was less expectation that the government should regulate **security** (5 participants, none holding the government solely responsible). Among those five, P32_A thought the government's duty was in *"setting guidelines, enforcing them."*

P7_A felt that a regulator's role was not about constant auditing but rather holding manufacturers responsible if they were to *"mess up"* with respect to security.

4.2.4 Shared Responsibility

Responsibility for **privacy** was often viewed as being shared by some combination of consumers, manufacturers, and government (21 participants). For instance, one participant thought both she and the manufacturer are obligated:

> *"I think I'm partially responsible in making sure that I don't put too much out there. But I think that the companies that control and own these, they need to make sure that people's information is not being put out there. Because at the end of the day, it affects us" (P37_A).*

Twenty-four thought responsibility for **security** was shared, mostly between user and manufacturer. A tech-savvy participant talked about this mutual obligation:

> *"If you have stronger security features that the device offers the user doesn't use, that's kind of the user's fault. If it doesn't offer certain level of security, that's the manufacturer's fault" (P10_A).*

We observed that participants perceived each actor (consumer, manufacturer, government) as having a role in filling in the gaps when other parties cannot or choose not to enact strong privacy and security measures. In the remainder of this section, we present the different combinations of responsible actors discussed by participants and how they viewed each actor as balancing the others.

Personal and Manufacturer. Most responses about shared responsibility for **security** were between device owners and manufacturers (19 participants), with much fewer (7) for **privacy**. From our analysis, we observe that the difference may be due to a recognition that both the device itself and the environment in which it is placed need to be secured, with only users themselves having the ability to secure the home network and set strong passwords on device companion apps. However, some acceptance of personal responsibility and mitigation implementation did not abdicate manufacturers, since there are aspects of security and privacy that users will never have control over (e.g., secure code, security of cloud services, protection of stored data and data in transit). Therefore, responsibility was often viewed as being shared, as expressed by a participant:

> *"I need to protect my passwords and things like that. But at the same time... you don't know what security features are built in, you don't know what any potential vulnerability might be. I think it's certainly a shared responsibility" (P24_A).*

As another example case, P1_A assumes personal responsibility both in purchase decision (*"It starts with us. We're bringing this device into our home"*) and by taking some simple mitigative actions (e.g., taping over cameras, not placing devices in more private areas of the home like bedroom). Yet, she also expects the manufacturer to do what she is not able to do with respect to managing data *"appropriately and securely"* and producing secure devices.

Given that smart home users may not know how to protect their devices and data, they look to manufacturers to provide them with more usable and transparent options. A smart home administrator commented about the need for better usability:

> *"I think the ability to control that data should be simpler than a multistep process, especially because the smart homes are very popular with people who don't know how to use technology" (P29_A).*

P3_A placed partial responsibility on herself for privacy (*"To the extent that you can do something about it, you should"*), but also felt the manufacturer should be more transparent:

> *"There's a certain responsibility to be transparent about what you're doing with people's data, protect personally-identifiable information, and to make it clear how you will use it up. I would want to know what their rules are about law enforcement, state access, and how they deal with data brokers and other companies."*

Even technology-savvy, advanced smart home users wanted manufacturers to fill in current gaps in available options. For example, when asked who he thinks is responsibility for the privacy of data collected by his smart home devices, P15_A commented: *"My personal perspective on it is that it's up to the user to be aware of what the device is doing and configure and use them appropriately according to your own needs."* However, he did not believe that consumers were given enough control:

> *"I think it would be ideal if the companies running the back end systems for these devices would give you either a little bit more control or be a lot more transparent about what they do with it and show themselves to be more responsible with that data."*

There is also a tension in that users do not always trust manufacturers' motives and ability to implement strong security, so they feel the need to take personal action. For example, P15_A viewed himself as being responsible in order to fill a gap left by manufacturers who fail to produce secure products:

> *"I'd like to see the vendors take more responsibility and take more action to secure their own devices. But because they don't always do that, and I don't always necessarily trust them to do that, I take it upon myself to be responsible for the security of these systems" (P15_A).*

Personal and Government. Only two participants thought that they and the government were responsible for **privacy** (none for **security**). One of those two, P31_A, discussed, *"We haven't even begun to really go down the road what the EU has as far as protecting privacy, but it's the government... and you personally, as much as you can to the extent practical."*

Manufacturer and Government. Nine participants thought manufacturers and government were jointly responsible for **privacy** but only three for **security**. Assignment of responsibility to the government or other regulatory bodies was usually rooted in response to lack of trust in manufacturers and belief that manufacturers were monetizing and selling smart home data. Government intervention was viewed as a standardizing construct that provides *"all the checks and balances" (P3_A)* on manufacturers so they do not circumvent privacy protections. For example, one participant commented:

> *"Voluntary consensus on privacy issues is almost impossible to get from the commercial sector... I think they need privacy guidelines at least from the government in order to adhere to them" (P13_A).*

Another participant claimed that companies are

> *"supposed to respect your privacy... If they fail,... next jurisdiction would be a government. The government has to watch them to make sure information is used for the right purposes" (P36_A).*

Personal, Manufacturer, and Government. Five participants viewed responsibility for **privacy** as being shared amongst themselves, manufacturers, and the government: *"It's the company... It's the government. But ultimately it's you and I" (P26_A).* Two participants viewed **security** as being shared among all three actors. A participant viewed privacy responsibility as being *"three-pronged... A third as a consumer, I should be aware, a third the company, and a third regulators and the government" (P25_A).* Another had a more in-depth explanation of his view of privacy responsibility:

> *"I think the company is responsible for it... in terms of government oversight, the government is in some way, shape, or form... Ultimately - and we're talking about accountability - you are responsible for your information because everyone else doesn't really care about you any more than you care about you" (P8_A).*

5 Discussion

In this section, we situate our results within prior literature on smart home privacy/security and IT responsibility. We then discuss the interdependent relationship between users, manufacturers, and third parties, and identify gaps and recommendations for how each actor can support the others.

5.1 Advancing Smart Home and Responsibility Research

In our study, we confirmed results of prior smart home studies indicating that well-known concepts in privacy and security translate into perceptions of smart home devices (cf. 2.1.1). As demonstrated in past studies [2, 47, 48], our research showed that users may have concerns, but they accept the risk in favor of perceived benefits. They choose to adopt privacy-violating technology and rarely take mitigative action, while accepting accountability for purchase and subsequent use. These behaviors reflect the privacy paradox [7]. This inaction may be due to several reasons. Users may have low security and privacy self-efficacy and experience security fatigue [54] and privacy resignation [42]. In addition, we found that taking action may be complicated due to hard-to-navigate configuration options or lack of any options at all (e.g., [34, 46]).

We advance research on responsibility by extending the investigation into the smart home domain, which has unique attributes as compared to traditional online and IT technology. For example, in our study, we observed that smart home devices are perceived as intrusive—always on and collecting sensitive data with ties to physical safety. Unfamiliarity with a new technology and the potential for many more devices in the home as compared to traditional IT devices adds complexity and vulnerability to the home network.

Similar to prior responsibility research (cf. 2.1.3, (especially [5]), we identified that users view smart home responsibility as being shared. We observed both active and passive responsibility, a perceived interdependent relationship, and, when necessary to motivate, a desire for a system of checks and balances for positive privacy/security outcomes. Although our participants felt that they bear some personal responsibility (as also discovered previously [5, 9, 19]), they often delegate responsibility to other entities (like manufacturers and government) when they do not feel equipped or incentivized to take action [27, 30, 50]. Tension may arise when users do not always trust the actors to whom they relegate responsibility, so they then look to others (government, industry oversight) to provide extra assurance [5]. Conversely, users may be resigned to having to take personal responsibility as a stopgap for lack of meaningful action on the part of manufacturers and government.

Moving beyond these similarities, we also identified differences from previous work. In prior smart home research (cf. 2.1.2), manufacturers and government are portrayed more as risks and bad actors [56, 66]. While some participants in our study did see these entities in potentially negative lights, they also recognized them as active partners in finding holistic solutions for smart home privacy and security. In addition, compared to prior findings that U.S. consumers rarely assign responsibility to their government for the protection of their digital assets [9, 19], we observed an appreciable number of our participants (roughly 37%) who thought government had responsibility for protecting smart home device privacy. This difference may be due to several potential reasons. First, the prior studies did not focus on smart home devices, rather connected devices in general, and may have lumped security, privacy, and safety together. Second, as compared to closed-ended survey choices, in our study, participants were able to organically assign responsibility in open-ended discussion. In addition, our study population was located in an area where the U.S. government is a major employer and more familiar.

Haney, Julie; Acar, Yasemin; Furman, Susanne M. "It's the Company, the Government, You and I: User Perceptions of Responsibility for Smart Home Privacy and Security." Paper presented at USENIX Security 2021, Vancouver, CA. August 11, 2021 - August 13, 2021.

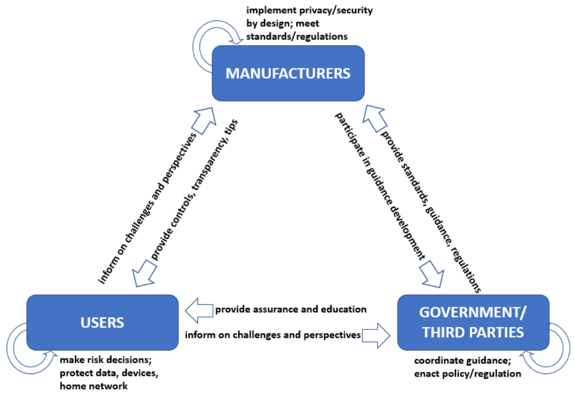

implement privacy/security by design; meet standards/regulations
MANUFACTURERS
inform on challenges and perspectives
provide controls, transparency, tips
participate in guidance development
provide standards, guidance, regulations
USERS
provide assurance and education
inform on challenges and perspectives
GOVERNMENT/ THIRD PARTIES
make risk decisions; protect data, devices, home network
coordinate guidance; enact policy/regulation

Figure 4: Perceived relationship between smart home users, manufacturers and third parties.

Progressing responsibility research into the smart home domain allows for identification of areas where users voiced the desire for immediate improvement (as described in the next section). The identification of perceived gaps is particularly valuable, given that this is a fledgling industry that currently lacks the maturity and full spectrum third-party support and guidance currently afforded to traditional IT.

5.2 Addressing Gaps

An overarching theme was the perceived interdependency between users, manufacturers, and government in a triad of responsibility. Through the eyes of smart home users, we observed disparities between the status quo and what consumers think should be happening. Disparities can point to future directions where researchers and practitioners should focus attention. As an example, if users accept responsibility but lack the ability to take action, discomfort with their smart home security and privacy may warrant action and investigation into how manufacturers can better support users or where third-party guidance or regulation may be beneficial.

In this section, we summarize problem areas and provide suggestions on how each actor can better be empowered to contribute to smart home security and privacy. The desired interdependent relationship identified by participants in our study is illustrated in Figure 4. Note that participants had a narrow view of oversight only coming from the government. However, recognizing that other, non-governmental organizations (e.g., non-profits, industry groups, standards organizations) may also be able to provide manufacturers and users with support, standards, and evaluations, we expand the government/regulatory actor into a broader third-party role. Our study also motivates future work related to each actor's potential contribution and needed support.

5.2.1 Problem Areas and Gaps

Users. We observed inconsistent relationships between being concerned, accepting personal responsibility, and taking privacy and security mitigative actions. Concerned participants did not always take action because of lack of knowledge, accepting trade-offs, and not valuing data collected by smart home devices (4.1). Those with privacy and security concerns did not always accept personal responsibility, and, sometimes, those who did not express concern still accepted responsibility (4.2.1).

There was also a marked disconnect between *feelings* of personal responsibility and *ability* to take active responsibility. While users may blame themselves for not actively protecting their security and privacy, they feel essentially powerless, resulting in a sense of privacy resignation and security fatigue. Most participants therefore believed that the privacy and security of their smart home should be a shared responsibility. Unfortunately, most of the burden is currently put on the user.

In order for users to be able to take *informed* personal responsibility, they need to better understand the risks, be given the opportunity to take action, and be educated about what steps they need to take. They also require reliable, objective information from manufacturers or trusted third parties to aid in purchase decisions. However, when researching smart home privacy and security, a minority relied on manufacturer-supplied information, with most participants trusting other online sources more.

Users who did not mention that they felt personally responsible mostly assigned responsibility to other actors, and not without reason. Concurrent research agrees that users' security and privacy needs in smart homes should go beyond what users can do (or are willing to do) and should be extensively supported by more powerful actors, like regulators and manufacturers (cf. Sections 2.1.2 and 2.2). This is complicated by users sometimes not trusting manufacturers or the government even when expecting support.

Manufacturers. Some participants believe manufacturers are competent with respect to privacy and security, often based on manufacturer reputation as opposed to transparent communication. Others doubt the willingness of manufacturers to implement strong privacy and security measures. They believe that manufacturers may not be incentivized to spend extra time/money on privacy and security for relatively inexpensive and disposable devices. Plus, added privacy restrictions may be counter to their business model of monetizing data, so participants believe that manufacturers may be purposely vague in what they reveal about data collection and use. Even though participants viewed manufacturers as being responsible, the reality is that some manufacturers may not know how to properly implement privacy and security, partly because many are new to developing smart products [25]. In addition, manufacturers may be unsure of what third-party guidance to follow since smart home privacy and security guidelines have not yet converged into widely agreed-upon standards.

The notion of manufacturers may also extend beyond those who develop smart home products. Third-party cloud and internet service providers and makers of the devices upon which smart home companion apps reside (e.g., smartphone and tablet manufacturers) may also hold some responsibility for security and privacy.

Government and Third Parties. While participants did not necessarily trust the government, they voiced a desire for third parties (including government) to develop smart home privacy and security regulation and guidelines to uphold and support manufacturer responsibility in a system of checks and balances. Participants were less understanding of how government guidance and regulation could help with security. This might be because participants were less clear about what security of smart home devices and data would mean for them.

While general privacy and security regulation is slowly being rolled out (e.g., CCPA and GDPR), few authoritative government regulations or guidelines for IoT/smart home privacy and security are available or widely adopted. Even though manufacturers sell devices globally, individual government organizations may create their own guidance or regulation that they want manufacturers to follow. (We note that none of the participants in this study lived in an area covered by any of the new privacy laws). In addition, industry groups may issue their own recommendations. Various guidelines from these organizations may or may not be consistent, which could result in manufacturer confusion on which to follow.

From a legal perspective, there is also debate on who should protect data and the boundaries of protection. Considering the newness of mandates in this area, legal constructs and interpretations will likely evolve.

5.2.2 Opportunities for Improvement

Based on identification of actions participants are willing/able to take and what they desire others to do, we offer the follow suggestions for strengthening the three-pronged, interdependent privacy/security relationship. We refer back to Results sections that inform our recommendations where appropriate.

What users can do. While manufacturers have a substantial responsibility to ensure smart home devices are privacy-respecting and secure, they cannot do everything and require users to be willing and active partners.

- ***Protection of data, devices, and home networks -*** Participants in our study thought they have some responsibility for configuring device options and setting strong passwords on device apps (4.1, 4.2.1). Recognizing that manufacturers have no control over the environment in which smart home devices are placed, users also need to protect their home networks, control device placement, and understand device capabilities and how those may impact or be used for privacy/security (4.1).
- ***Due diligence in understanding and accepting risks -*** Smart home users make privacy and security tradeoffs (4.1). Although they should be better supported in making these decisions and understanding risks, they are ultimately responsible for making informed decisions in line with their own privacy and security expectations and needs (4.2.1).

What third parties can do. Third parties, including oversight, government, and consumer-focused organizations, can provide support and guidance for smart home users and manufacturers. Users seem receptive to some government oversight and outside guidance for manufacturers, especially in the privacy area (4.2.3).

- ***Oversight and development of standards and guidelines for smart home privacy and security -*** Government bodies can protect consumers' privacy and security and aid manufacturers by issuing voluntary guidance or regulations when appropriate on recommended privacy and security implementations and options (e.g., [22,23]). Non-profits, industry forums, standards organization, etc. can also contribute to building a more universal consensus of what constitutes minimum privacy and security measures in smart home devices, for example via baselines [14,40] and product labels/ certifications [15,21,59]. Because users often lack the knowledge to take action on their own (4.1), recommendations should take user considerations into account, for example, with suggestions on how manufacturers might consider user limitations throughout the entire product lifecycle [23].
- ***Consumer education -*** Third parties can provide resources that educate users on smart home privacy and security issues and provide actionable configuration tips (4.1).

What manufacturers can do. Because smart home users may not be technology- or security-savvy (4.1), we found that users often want to rely on manufacturers (4.2.2) to fill this gap in several ways:

- ***Usable privacy/security interfaces -*** Provide an interface that makes it easy for users to configure privacy/security options (e.g., opt in/out), while not overburdening users with too many options.
- ***Transparent privacy and security practices -*** Be more forthcoming about what privacy and security options are available, which features are built into the products, and options/features that are not available but may be expected. To address user's distrust of manufacturer motives (4.2.2), make this information easier for consumers to find (e.g., on vendor websites or device help/support screens). Also provide more readable and accessible privacy policies that transparently communicate how data is collected, stored, and used.

- ***Privacy and security by design -*** Alleviate user burden of having to configure extra privacy and security options (4.1) by making an honest effort to provide strong "out-of-the-box" privacy and security features. Care should be taken, however, to ensure these features do not impact usability. Follow privacy/security guidance provided by reputable third parties, for example, practicing data minimization principles by only collecting data that is required to fulfill functionality and not violating contextual integrity (e.g., Alexa transmitting audio to find answers, but not storing voice recordings).

- ***Standards and guidance participation -*** In conjunction with our participants' desire for third parties to develop privacy/security guidance and standards (4.2.3), manufacturers should actively engage in coming to consensus on minimum smart home privacy/security recommendations. These recommendations can then be used in evaluations that contribute to product labels and certifications.

- ***Consumer education -*** Via app interfaces and help/support documentation, give consumers objective tips on how to best configure their devices with privacy/security in mind to account for users' uncertainty on what to do and how to do it (4.1).

5.2.3 Research Opportunities

Our exploratory study motivates future research direction into product labels, privacy/security education and communication efforts for users and smart home device manufacturers, interface design for configuring privacy and security features, and suggested standards for smart home privacy/security. There may also be value in more exploration into who should be responsible for implementing these improvements as well as receptivity and ability to take on additional duties. For example, little research has been done to capture the smart home manufacturer perspective. As such, future research may be warranted to determine where manufacturers are most challenged and how to best provide support and value. The practicalities of manufacturers implementing our proposed security/privacy recommendations also need to be better understood, (e.g., whether certain features can be implemented on devices with limited memory and processing power). Exploration of appropriate incentives that might frame the production of secure and private devices as a competitive advantage would also be valuable. We acknowledge that responsibility perceptions may be influenced by cultural, national, and political factors, so there is a need for extending current research into broader populations, including those outside the U.S. We also see an opportunity for increased real-world transfer of the knowledge gained from user-centered research efforts in this area to inform manufacturers and guideline developers. This study has already informed some of the user-centric considerations in NIST security guidance for manufacturers [23].

6 Conclusion

In a qualitative research study of 40 smart home users, we expand the discourse on smart home security and privacy by investigating where users perceive responsibility for their smart home security and privacy. We find a theme of an interdependent relationship in which participants assume some personal responsibility but also assign responsibility to manufacturers and government/third parties when they cannot or are not willing to mitigate their concerns. We identify areas needing improvement in the current smart home privacy and security domain and distill how actors can take steps to fill these gaps. Achieving a more balanced relationship may take some of the burden off of users and provide better support to manufacturers, leading to less vulnerable systems and greater adoption of smart home technologies.

Acknowledgements

We would like to thank the anonymous reviewers, our shepherd Marshini Chetty, and our colleagues Sascha Fahl, Adam Aviv, Michael Fagan, Kevin Mangold, and Brian Stanton for their helpful comments on drafts of this paper. We would also like to thank Mary Theofanos for her input during initial study design.

References

[1] Noura Abdi, Kopo M Ramokapane, and Jose M Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Symposium on Usable Privacy and Security*. USENIX, 2019.

[2] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[3] Alessandro Acquisti, Leslie K John, and George Loewenstein. What is privacy worth? *The Journal of Legal Studies*, 42(2):249–274, 2013.

[4] Noah Apthorpe, Sarah Varghese, and Nick Feamster. Evaluating the contextual integrity of privacy regulation: Parents' IoT toy privacy norms versus COPPA. In *USENIX Security Symposium*, pages 123–140, 2019.

[5] Soumava Bandyopadhyay. Antecedents and consequences of consumers online privacy concerns. *Journal of Business & Economics Research*, 7(3), 2009.

[6] Rosaline S. Barbour. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *British Medical Journal*, 322(7294):1115–1117, 2001.

[7] Susanne Barth and Menno D.T. de Jong. The privacy paradox – investigating discrepancies between expressed privacy concerns and actual online behavior – a systematic literature review. *Telematics and Informatics*, 34(7):1038 – 1058, 2017.

[8] Berg Insight. Smart homes and home automation. http://www.berginsight.com/ReportPDF/ProductSheet/bi-sh7-ps.pdf, 2019.

[9] Jen Caltrider. 10 fascinating things we learned when we asked the world 'how connected are you?'. https://blog.mozilla.org/blog/2017/11/01/10-fascinating-things-we-learned-when-we-asked-the-world-how-connected-are-you/, 2017.

[10] George Chalhoub, Ivan Flechais, Norbert Nthala, and Ruba Abu-Salma. Innovation inaction or in action? the role of user experience in the security and privacy design of smart home cameras. In *Symposium on Usable Privacy and Security*, pages 185–204. USENIX, 2020.

[11] Chola Chhetri and Vivian Genaro Motti. Eliciting privacy concerns for smart home devices from a user centered perspective. In *International Conference on Information*, pages 91–101. Springer, 2019.

[12] Jessica Colnago, Yuanyuan Feng, Tharangini Palanivel, Sarah Pearman, Megan Ung, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. Informing the design of a personalized privacy assistant for the internet of things. In *CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM, 2020.

[13] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA, 4th edition, 2015.

[14] Council to Secure the Digital Economy. The C2 consensus on IoT security baseline capabilities. https://securingdigitaleconomy.org/projects/c2-consensus/, 2019.

[15] CTIA Certification. CTIA certification resources. https://www.ctia.org/about-ctia/programs/certification-resources, 2020.

[16] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We value your privacy... now take some cookies: Measuring the GDPR's impact on web privacy. *arXiv preprint arXiv:1808.05096*, 2018.

[17] Department for Digital, Culture, Media and Sport. Code of practice for consumer IoT security. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/773867/Code_of_Practice_for_Consumer_IoT_Security-_October_2018.pdf, 2018.

[18] Departments of Commerce and Homeland Security. A report to the president on enhancing the resilience of the internet and communications ecosystem against botnets and other automated, distributed threats. https://csrc.nist.gov/CSRC/media/Publications/white-paper/2018/05/30/enhancing-resilience-against-botnets--report-to-the-president/final/documents/eo_13800_botnet_report_-_finalv2.pdf, May 2018.

[19] Leyla Dogruel and Sven Joeckel. Risk perception and privacy regulation preferences from a cross-cultural perspective: A qualitative study among German and US smartphone users. *International Journal of Communication*, 13:20, 2019.

[20] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an IoT privacy and security label? In *IEEE Symposium on Security and Privacy*, 2020.

[21] Pardis Emami-Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Faith Cranor. Exploring how privacy and security factor into IoT device purchase behavior. In *CHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[22] ENISA. Good practices for security of IoT - Secure software development lifecycle. https://www.enisa.europa.eu/publications/good-practices-for-security-of-iot-1, 2019.

[23] Michael Fagan, Katerina N. Megas, Karen Scarfone, and Matthew Smith. NISTIR 8259 Foundational cybersecurity activities for IoT device manufacturers. https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8259.pdf, 2020.

[24] Michael Fagan, Mary Yang, Allen Tan, Lora Randolph, and Karen Scarfone. Draft NISTIR 8267 Security review of consumer home Internet of Things (IoT) products. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8267-draft.pdf, 2019.

[25] Federal Trade Commission. Internet of things privacy and security in a connected world. https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf, 2015.

[26] Kevin Fu, Tadayoshi Kohno, Daniel Lopresti, Elizabeth Mynatt, Klara Nahrstedt, Shwetak Patel, Debra Richardson, and Ben Zorn. Safety, security, and privacy threats posed by accelerating trends in the internet of things. Technical report, Computing Community Consortium Report 29, no. 3, 2017.

[27] Susanne Furman, Mary Frances Theofanos, Yee-Yin Choong, and Brian Stanton. Basing cybersecurity training on user perceptions. *IEEE Security & Privacy*, 10(2):40–49, 2011.

[28] Christine Geeng and Franziska Roesner. Who's in control?: Interactions in multi-user smart homes. In *CHI Conference on Human Factors in Computing Systems*, page 268. ACM, 2019.

[29] GfK. Future of smart home study global report. https://www.gfk.com/fileadmin/user_upload/dyna_content/GB/documents/Innovation_event/GfK_Future_of_Smart_Home__Global_.pdf, 2016.

[30] Joshua B. Gross and Mary Beth Rosson. Looking for trouble: understanding end-user security management. In *Symposium on Computer Human interaction for the Management of Information Technology*, pages 10–es, 2007.

[31] GutCheck. Smart home device adoption. https://resource.gutcheckit.com/smart-home-device-adoption-au-ty, 2018.

[32] Julie M. Haney, Susanne M. Furman, and Yasemin Acar. Smart home security and privacy mitigations: Consumer perceptions, practices, and challenges. In *International Conference on Human-Computer Interaction*, 2020.

[33] Harris Interactive. Consumer internet of things security labelling survey research findings. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/798543/Harris_Interactive_Consumer_IoT_Security_-Labelling_Survey_Report.pdf, 2019.

[34] Woodrow Hartzog. Website design as contract. *Am. UL Rev.*, 60:1635, 2010.

[35] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. Rethinking access control and authentication for the home internet of things (IoT). In *USENIX Security Symposium*, pages 255–272, 2018.

[36] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Workshop on New Security Paradigms*, pages 133–144, 2009.

[37] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks. In *CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. ACM.

[38] Internet of Things Privacy Forum. Clearly opaque: Privacy risks of the IoT. https://www.iotprivacyforum.org/research/, 2018.

[39] Iulia Ion, Rob Reeder, and Sunny Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *Symposium On Usable Privacy and Security*, pages 327–346. USENIX, 2015.

[40] IoT Security Foundation. Secure design best practice guides. https://www.iotsecurityfoundation.org/wp-content/uploads/2019/11/Best-Practice-Guides-Release-2.pdf, 2019.

[41] Carlos Jensen, Colin Potts, and Christian Jensen. Privacy practices of internet users: self-reports versus observed behavior. *International Journal of Human-Computer Studies*, 63(1-2):203–227, 2005.

[42] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. In *ACM on Human-Computer Interaction*. ACM, 2018.

[43] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Privacy Enhancing Technologies*, 2019(4):250–271, 2019.

[44] Shrirang Mare, Logan Girvin, Franziska Roesner, and Tadayoshi Kohno. Consumer smart homes: Where we are and where we need to go. In *International Workshop on Mobile Computing Systems and Applications*, pages 117–122, 2019.

[45] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. In *ACM on Human-Computer Interaction*, page 72, 2019.

[46] Craig RM McKenzie, Michael J Liersch, and Stacey R Finkelstein. Recommendations implicit in policy defaults. *Psychological Science*, 17(5):414–420, 2006.

[47] Patricia A Norberg, Daniel R Horne, and David A Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 41(1):100–126, 2007.

[48] Chanda Phelan, Cliff Lampe, and Paul Resnick. It's creepy, but it doesn't bother me. In *CHI Conference on Human Factors in Computing Systems*, page 5240–5251, New York, NY, USA, 2016. ACM.

[49] PwC. Smart home, seamless life. https://www.pwc.fr/fr/assets/files/pdf/2017/01/pwc-consumer-intelligence-series-iot-connected-home.pdf, January 2017.

[50] Karen Renaud, Melanie Volkamer, and Arne Renkema-Padmos. Why doesn't Jane protect her privacy? In *International Symposium on Privacy Enhancing Technologies*, pages 244–262, 2014.

[51] Angela Sanguinetti, Beth Karlin, and Rebecca Ford. Understanding the path to smart home adoption: Segmenting and describing consumers across the innovation-decision process. *Energy research & Social Science*, pages 274–283, 2018.

[52] Irina Shklovski, Scott D Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *CHI Conference on Human Factors in Computing Systems*, pages 2347–2356. ACM, 2014.

[53] The Internet Society. Securing the internet of things: A Canadian multistakeholder process draft report. https://iotsecurity2018.ca/wp-content/uploads/2019/02/Enhancing-IoT-Security-Draft-Outcomes-Report.pdf, 2019.

[54] Brian Stanton, Mary F. Theofanos, Sandra Spickard Prettyman, and Susanne Furman. Security fatigue. *IT Professional*, 18(5):26–32, 2016.

[55] State of California. SB-327 Information privacy: connected devices. https://leginfo.legislature.ca.gov, September 2018.

[56] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. "I don't own the data": End user perceptions of smart home device data practices and risks. In *Symposium on Usable Privacy and Security*. USENIX, 2019.

[57] Madiha Tabassum, Jess Kropczynski, Pamela Wisniewski, and Heather Richter Lipford. Smart home beyond the home: A case for community-based access control. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, 2020.

[58] Mary Theofanos, Brian Stanton, Susanne Furman, Sandra Spickard Prettyman, and Simson Garfinkel. Be prepared: How US Government experts think about cybersecurity. In *Workshop on Usable Security*, USEC '17, pages 1–11, 2017.

[59] UL. IoT security rating. https://ims.ul.com/IoT-security-rating, 2020.

[60] European Union. General data protection regulation. http://data.europa.eu/eli/reg/2016/679/oj, 2016.

[61] Blase Ur, Jaeyeon Jung, and Stuart Schechter. The current state of access control for smart devices in homes. In *Workshop on Home Usable Privacy and Security*, volume 29, pages 209–218, 2013.

[62] Rick Wash and Emilee Rader. Too much knowledge? Security beliefs and protective behaviors among United States internet users. In *Symposium On Usable Privacy and Security*, pages 309–325, 2015.

[63] Ryan West, Christopher Mayhorn, Jefferson Hardee, and Jeremy Mendel. *Social and Human Elements of Information Security: Emerging Trends and Countermeasures*, chapter The weakest link: A psychological perspective on why users make poor security decisions, pages 43–60. IGI Global, 1 edition, 2009.

[64] Meredydd Williams, Jason RC Nurse, and Sadie Creese. Privacy is the boring bit: User perceptions and behaviour in the internet-of-things. In *Conference on Privacy, Security and Trust*, pages 181–18109. IEEE, 2017.

[65] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM, 2019.

[66] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *Symposium on Usable Privacy and Security*, 2017.

[67] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home IoT privacy. *ACM on Human-Computer Interaction*, 2.

Haney, Julie; Acar, Yasemin; Furman, Susanne M. "It's the Company, the Government, You and I: User Perceptions of Responsibility for Smart Home Privacy and Security." Paper presented at USENIX Security 2021, Vancouver, CA. August 11, 2021 - August 13, 2021.

A Participant Demographics

ID	Gen	Age	Ed	Occupation	Device Type				
					Sec	Ent	Env	Appl	Asst
P1_A	F	50-59	M	Liaison	X		X		X
P2_A	M	30-39	M	Lead engineer	X	X	X		X
P3_A	F	40-49	M	Professor	X	X	X	X	X
P4_A	M	60+	M	Retired	X	X			
P6_U	F	30-39	B	Events manager	X	X	X	X	X
P7_A	M	30-39	B	Software engineer	X	X	X	X	X
P8_A	M	30-39	B	Federal employee	X	X	X	X	X
P9_A	F	30-39	M	Educationist	X	X	X		X
P10_A	M	30-39	B	Computer scientist	X	X	X	X	X
P11_A	M	50-59	M	Electrical engineer	X	X	X		X
P12_U	F	30-39	M	Administrative assistant	X	X	X		X
P13_A	M	50-59	M	Manager, cognitive scientist	X	X	X	X	X
P14_U	F	40-49	H	Information specialist	X	X	X		X
P15_A	M	30-39	B	Computer scientist	X	X	X		
P16_A	M	40-49	M	Research chief	X	X	X		X
P17_A	F	30-39	M	Systems engineer	X	X	X	X	X
P18_A	M	30-39	B	Business consultant	X	X	X		X
P19_A	M	50-59	B	Retail services specialist	X	X	X	X	X
P20_A	F	30-39	B	Administrator		X			
P21_U	F	18-29	B	Human resources manager	X	X	X	X	X
P22_A	M	30-39	B	Executive admin assistant	X	X	X	X	X
P23_A	F	40-49	M	Community arts specialist	X	X	X		X
P24_A	M	40-49	B	Operational safety analyst		X	X		X
P25_A	M	30-39	B	Program management analyst	X	X	X	X	X
P26_A	M	30-39	B	Analyst	X	X	X		X
P27_A	F	40-49	M	Program coordinator	X	X	X	X	X
P28_A	F	50-59	B	Consultant	X		X		X
P29_A	M	18-29	M	Events coordinator	X	X	X		X
P30_U	F	18-29	B	Event planner	X	X	X		X
P31_A	F	30-39	M	Lobbyist	X	X	X		X
P32_A	M	30-39	B	Health educator		X	X	X	X
P33_A	M	18-29	B	Senior technology analyst	X	X	X		X
P34_A	M	40-49	B	Financial analyst	X	X	X	X	X
P35_A	M	40-49	M	Accountant	X	X	X	X	X
P36_A	F	30-39	B	Project manager	X	X	X		X
P37_A	F	40-49	M	Assistant principal	X	X	X		
P38_U	F	60+	M	Special educator		X	X		X
P39_U	M	60+	M	Retired		X	X		X
P40_U	F	30-39	C	Customer service rep	X	X	X		X
P41_A	M	40-49	B	Security	X	X	X		X
				Total	35	38	38	15	36

Table 1: Participant Demographics. ID: A - smart home administrators/installers, U - smart home users; Gen (Gender); Ed (Education): M - Master's degree, B - Bachelor's degree, C - some college, H - High school; Device Type: Sec - Home security, Ent - Home entertainment, Env - Home environment, Appl - Smart appliance, Asst - Virtual assistant. Interviewed couples: P6_U and P7_A, P29_A and P30_U, P38_U and P39_U, P40_U and P41_A.

B Interview Questions

NOTE: Questions yielding responses that contributed to this paper are in **bold**.

SECTION A: TERMINOLOGY

1. You may have heard the term “internet of things,” or IoT for short. Can you talk a little about what you think the internet of things is?
2. You may have heard the term “smart devices.” What about devices makes them “smart?”
3. What does it mean to have a smart home?
4. What do you think is the relationship, if any, between the internet of things and smart devices?

SECTION B: PURCHASE & GENERAL USE
[Review list of smart home devices before beginning this section.]

5. Who was involved in the decision to purchase the smart home devices?
6. **What are the reasons the smart home devices were purchased?**
   - **How did you (or a household member) learn about the devices before buying them?**
7. **What hesitations, if any, did you have about getting the devices prior to purchase?**
8. For what purposes do you use your smart home devices?
9. How do you access the devices – remotely with an app, while physically in the home, or both?
   - *If using a virtual assistant:* How do you access your devices using [insert assistant name]?
   - *If using a hub:* Do you use the hub app to access your devices, or do you use an individual app specific to each device?
10. How do others in your household use the smart home devices?
11. What do you like most about the devices? What are the benefits, if any, of having these devices?
12. **What do you like least or dislike about the devices?**
13. How have your opinions or expectations of the devices changed, if at all, from the time you first used them until now?
14. **What concerns, if any, do you have about the devices?**
15. **In what ways, if any, have you changed your behaviors because of your smart home devices?**
16. **In what ways, if any, have you become reliant on your smart home devices?**
17. What do the other members of your household think about the smart home devices?
18. Have you had visitors to the home who have had to use the smart home devices?
    - *If yes:* How did they use the devices? What did they think?
19. What smart home devices, if any, have you had in the past, but are no longer using?
    - **What are the reasons for no longer using this device?**
20. **What kinds of things would you like to be able to do with your devices, but haven’t, don’t know how, or are not sure that you can?**
21. What devices would you like to get in the future? For what reasons?

SECTION C: INSTALLATION/TROUBLESHOOTING

22. Who installed the smart home devices?
23. Who administers (configures or maintains) the smart home devices?

For Installers:

24. In general, what was your experience with the installation of the devices?
    - What went well?
    - What didn’t go as well?
25. Have you ever had to reinstall a device? If so, what were the reasons for the reinstallation?
26. *If have more than one device:* What has been your experience adding additional devices to the home?

For DIYers:

27. In the screening questionnaire you indicated you build your own or create extensions for your smart home devices and platforms. Can you briefly summarize what you’ve done?

For Administrators:

28. What configuration changes, if any, have you made to the devices since installation?

- *If participant makes configuration changes:* How often do you make changes?

29. How do you know that updates are available or needed?

30. How are updates done on your device - automatically or do you have to initiate them?

    - *If manual initiation:* How often do you check for updates?
    - How do you decide whether to update or not update?

For Everyone:

31. How do you try to figure out how to do something new with your devices?

    - What sources do you consult or use?
    - *If have a voice assistant:* What has been your experience, if any, adding new skills to your voice assistant?

32. What kinds of problems, if any, have you encountered while using your smart home devices?

    - How did you go about trying to resolve those problems?

SECTION D: PRIVACY

33. **What type of information, if any, do you think the devices are collecting?**

    - **Which of this information, if any, would you consider to be personal?**

34. **Where do you think the information goes?**

35. **In what ways, if any, does your device or the device manufacturer provide a means to control or manage what information is collected and how it is shared?**

36. **What are your concerns, if any, about how information is collected, stored, and used and who can see that information?**

    - **In what ways, if any, have you acted to minimize or alleviate some of those concerns?**
    - **What kinds of actions would you like to be able to take to address your concerns, but haven't, don't know how, or are not sure that you can?**

37. **Who do you think is responsible for protecting the privacy of information collected by your smart home device?**

SECTION E: SECURITY

38. **What are your concerns, if any, about the security of your devices?**

    - **In what ways, if any, have you acted to minimize or alleviate some of those concerns?**
    - **What kinds of actions would you like to be able to take to address your concerns, but haven't, don't know how, or are not sure that you can?**

39. What restrictions, if any, are placed on who in your home can use the devices and what they can do?

40. How do you authenticate to or get into any apps associated with the device?

    - What issues or problems, if any, have you experienced with authentication?

41. Does more than one person in your household use an app to access the same device?

    - Does more than one person use the same account and authentication to access the app?
    - What concerns, if any, do you have with multiple people having access to the app?

42. **Who do you think is responsible for the security of your smart home devices?**

SECTION F: SAFETY

43. In what ways, if any, do you think the devices contribute to safety?

44. **In what ways, if any, do you think the devices might pose a safety risk?**

SECTION G: CONCLUSION

45. Is there anything else you'd like to add related to anything we've talked about?

C Codebook for Privacy & Security Concepts

Code	Description
Privacy	Related to (potentially sensitive) data collected by smart home devices
Attitudes	When a participant talks about their mindset or point of view with respect to privacy
Concerns	When a participant talks about some worry they have about how their smart home data is collected, used, and protected
Audio/video recording	Exposure of data from devices that record
Data breach	Data being stolen from manufacturers
Financial	Financial information and accounts being exposed by devices/apps
Government access	Data being exposed or devices being hacked by domestic or foreign nation states
Habits/profiles	Data or the aggregation of data revealing information about the household and occupants
Lack of concern	Evidence of participants not having worries about smart home device privacy/data
Selling data/ads	Manufacturers selling data to third parties and/or data being used for targeted advertising
Unknowns	Discomfort with not knowing what data is being collected by devices, who has access to the data, or how the data is used
Information collection/use	When a participant talks about the collection and use of their smart home data
Control options	Mentions of options or lack thereof for managing data collection and use
Info collection	The types of data devices are collecting
Info destination	Where data goes
Info use	How data is used
Personal information	Which data is considered to be personal
Privacy policies/EULAs	Perceptions of and experiences with privacy information provided by device manufacturers
Mitigations	How participants tried (or didn't try) to lessen the privacy impact of their smart home devices or alleviate their privacy concerns
Access control	Limiting who/what devices can connect to or use the smart home devices/apps
Authentication	Setting passwords or biometric authentication for smart home devices/apps
Choosing devices that protect privacy	Taking privacy into account when selecting devices for purchase/use
Configuring options	Setting available privacy-related controls on the device
Lack of mitigations	How/why participants do not implement privacy-related mitigations
Limiting audio/visual	Restricting or being careful about what audio and video is recorded by the devices
Monitoring data	Inspecting network traffic entering/leaving the home network
Network configuration	Settings and actions taken on the home network, including Wi-Fi
Updates	Implementing patches or upgrading products/versions
Responsibility	When a participant talks about who is accountable for the privacy of data collected by smart home devices
Manufacturer	Responsibility assigned to the makers of smart home devices
Personal	Responsibility assigned to the participants themselves or smart home owners in general
Shared	Responsibility assigned to more than one actor/entity
Third-party	Responsibility assigned to governments or other non-manufacturer entities/organizations
Wishlist	What participants would like to have to increase the privacy of data collected by their smart home devices

Code	Description
Security	Related to how smart home devices and data are protected against unauthorized use or modification
Attitudes	When a participant talks about their mindset or point of view with respect to security
Authentication	Related to how participants establish or validate their identity when using smart home devices
Concerns	When a participant talks about some worry they have about the security of their smart home devices/data
Audio/video recording	Unauthorized exposure of data from devices that record voices or images
Data breach	Data being stolen from manufacturers
Exploitation of devices	Devices being hacked
Financial	Financial information and accounts being exposed by devices/apps
Government access	Data being exposed or devices being hacked by domestic or foreign nation states
Lack of concern	Evidence of participants not having worries about device security
Linkage to other accounts	Compromise of one account/device can lead to compromise of other accounts
Lack of out-of-the-box security	Devices are not secure when purchased
Physical security/safety	Related to potential harm of occupants or household items
Updates	Issues caused by installing patches or upgrading devices
Wi-Fi access	Unauthorized access to devices leading to compromise of Wi-Fi network
Mitigations	How participants tried (or didn't try) to alleviate their security concerns or improve security of their devices
Access control	Limiting who/what devices can connect to or use the smart home devices/apps
Authentication	Setting passwords or biometric authentication for smart home devices/apps
Choosing secure devices	Taking security features into account when selecting devices for purchase/use
Configuring options	Setting available security-related controls on the device
Lack of mitigations	How/why participants do not implement security-related mitigations
Limiting audio/visual	Restricting or being careful about what audio and video is recorded by the devices
Limiting information	Only sharing necessary information with device apps or using false information
Network configuration	Settings and actions taken on the home network, including Wi-Fi
Physical security	Actions that protect the physical device or prevent physical tampering
Updates	Implementing patches or upgrading products/versions
Responsibility	When a participant talks about who is accountable for the security of their smart home devices
Manufacturer	Responsibility assigned to the makers of smart home devices
Personal	Responsibility assigned to the participants themselves or smart home owners in general
Shared	Responsibility assigned to more than one actor/entity
Third-party	Responsibility assigned to governments or other non-manufacturer entities/organizations
Wishlist	What participants would like to have to increase the security of smart home devices/data
Household member disparities	Differences or inequalities in smart home security/privacy knowledge among those living in a house
Sense of control/agency	Perceptions of feeling in charge of their smart home devices
Trust	Perceptions of confidence in the smart home devices or the manufacturers to protect their privacy and security

NIST SP 1285
September 2022

High temperature reflectometer for spatially resolved spectral directional emissivity of laser powder bed fusion processes

David C. Deisenroth[1,a], Leonard Hanssen[a], and Sergey Mekhontsev[b]
[a]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899
[b]Jung Research and Development Corporation, 5508 Cromwell Drive, Bethesda, MD 20816

ABSTRACT

Additive manufacturing involving layer-wise selective laser melting of a powder material, or laser powder bed fusion (LPBF), is a fast-growing industry. At the Additive Manufacturing Metrology Testbed (AMMT) at the United States National Institute of Standards and Technology (NIST) an integrating hemispherical reflectometer has recently been developed to facilitate measurements of spatially resolved reflectance of the laser-melting heat affected zone (HAZ) during the LPBF process. Reflectance is then used to determine spatially resolved emissivity. The design features of the hemispherical-directional reflectometer are discussed. Then, the reflectometer performance and measurement uncertainties are detailed. A two-dimensional map of emissivity and emissivity uncertainty of the HAZ around a meltpool of high-purity nickel are presented. It is found that emissivity measurements are in good agreement with literature values at the melting point of high-purity nickel with acceptable uncertainty.

Keywords: High-temperature baffle-less reflectometer, emissivity of high-temperature metals, laser powder bed fusion

1. INTRODUCTION

Laser powder bed fusion (LPBF) is a layer-wise selective melting of the metal powder by scanning a high-power laser across the powder bed, which is located inside a process chamber with low oxygen content and laminar flow of protective gas. It is a fast-growing industry with unique fabrication capabilities compared to more traditional casting and forging [1]. Physical understanding of the processes taking place in laser-based additive manufacturing processes, such as LBPF and others, can be significantly enhanced by the knowledge of the thermodynamic surface temperature distribution [2]. The utility of thermodynamic surface temperature distribution includes, but is not limited to, the study of the rapid solidification of molten metal, which defines the metallographic structure—and associated mechanical properties—of the resulting part.

To establish traceable radiance-based temperature measurements we have selected the only first principle approach which is applicable, (1) direct measurement of spectral radiance of the heat-affected zone (HAZ) by comparison with a radiance standard, (2) indirect measurement of spectral emissivity, by illuminating the sample with a uniform hemispherical source and comparing the reflected radiant flux to that from a calibrated standard, and (3) calculation of the surface temperature distribution. This general approach has been realized and validated previously at NIST [3], and here is applied to the LPBF process, which is quite different from the one normally encountered in laboratory reflectometers. The scope of this paper is limited to the reflectometer design, performance of emissivity measurements, and evaluation of uncertainties.

The relatively small dimensions of the HAZ (with typical size of molten metal area less than 0.3 mm × 1.2 mm) and the dynamics of the process necessitate retrieval of spatially distributed data at a high spatial resolution. This comes at a cost of using a much more complicated sensor with relatively large imperfections and uncertainties as compared with a single element optical detector normally used in optical metrology [4]. In terms of requirement for reflectometry, this means using hemispherical illumination and directional (conical but within a small collection angle) imaging.

The temperatures encountered in a typical LPBF process range from room temperature to 3000 °C and above, with a special interest at the temperatures of the solidification and crystallographic phase changes (500 °C and above). Since the probing light of the reflectometer must compete with the self-emitted light from the hot scene, and in combination with

[1] david.deisenroth@nist.gov; phone 1 301 975-2594

Deisenroth, David; Hanssen, Leonard; Mekhontsev, Sergey. "High temperature reflectometer for spatially resolved spectral directional emissivity of laser powder bed fusion processes." Paper presented at Reflection, Scattering, and Diffraction from Surfaces VII, San Diego, CA, US. August 24, 2020 - August 28, 2020.

the hemispherical-directional geometry (as mentioned above), this calls for an unusually high-power light sources in the reflectometer.

A further complication arises from the dynamic character of the process, in which the laser spot is scanned across the build area. Every frame of the imager is registered at a different location, which further complicates reflectance measurements due to sample spatial nonuniformity and possible defocusing of the laser. Since use of a stationary field of view will result in motion blur, it is customary to use imaging systems which share the scanning system with the process laser. This allows for avoidance of some problems with relative motion of the process and the field of view, but brings a new set of problems due to changes in focusing and optical path length for the wavelengths of the laser and reflectometry wavelength, as well as a lower collection efficiency due to the optics which are optimized for the laser wavelength to avoid damage by the high power laser. This results in additional uncertainties due to position-dependent defocusing of the laser and image.

Finally, there are some additional aspects of the process due to the presence of the process by-products (such as hot and condensed metal vapor), which require a directional flow of a shield gas which can significantly affect the process [5]. This introduces new sources of uncertainty, as well as further complicating the reflectometer design. Some of these effects have yet to be evaluated in detail, which is planned for a later date.

As stated previously, the scope of this paper is limited to the reflectometer design, the performance of emissivity measurements, and uncertainty evaluation. Section 2 describes the measurement approach, including the measurement equation, reflectometer design, and signal corrections for directional imaging. Section 3 details the uncertainty evaluation of test conditions, reflectometer illumination, and directional imaging. Finally, Section 4 describes measurement of a two-dimensional map of emissivity and emissivity uncertainty of the HAZ around a meltpool of high-purity nickel, followed with conclusions in Section 5.

2. EXPERIMENTAL SETUP AND MEASUREMENT APPROACH

The experiments were performed in the NIST Additive Manufacturing Metrology Testbed (AMMT) [6,7]. The AMMT is a custom LPBF research platform that is designed to be highly configurable for characterization of all aspects of the LPBF process. The AMMT includes a removable carriage that contains the build-well and a large metrology-well, both of which may be moved laterally within the large build chamber. The laser is an Yb-doped continuous wave (CW) fiber laser with an emission wavelength of 1070 nm. The delivered laser power can be adjusted from 20 W to about 385 W, with a 4σ diameter (D4σ, representing diameter within which about 95% of the Gaussian laser power profile is contained) spot size that is adjustable from 45 µm to more than 200 µm. The laser spot can be scanned with full control of the laser scan path/strategy at 100 kHz and laser power control at 50 kHz, with scan velocity of more than 4000 mm/s.

The core elements of the AMMT meltpool thermographic systems include (1) a high power fiber laser system emitting at 1070 nm; (2) an optical scanner (galvanometer), used to direct process laser spot; (3) beam splitter and optical components enabling co-axial meltpool imaging configuration; and (4) the sample under study, which can be accurately positioned and aligned with the object plane of the co-axial laser/imager optical path and is surrounded by an environmental enclosure with a shield gas flow. Additional thermographic equipment, which is referred to as the TEMPS system (Temperature and Emissivity of Melts, Powders, and Solids), includes an external (coaxial) imaging system, which is optically combined with the process laser and is used to measure the radiance distribution across the HAZ of the process.

An “indirect” method of emissivity measurement is employed in this work, in which the laser melting process is uniformly illuminated with uniform radiance across the hemisphere by a hemispherical reflectometer. The reflectometer uses a hemispherical-directional geometry, in which a ring of light emitting diodes (LEDs) around the equator of the hemisphere is optically integrated within the reflectometer to provide the uniform illumination. The laser melting scene is then imaged directionally through imaging optics that are coaxial with the heating laser, allowing for stationary viewing of the meltpool relative to the laser heating location. The measurement is performed once with the LED illumination on, and once with the LED illumination off. This approach facilitates spatially resolved radiance and reflectance/emissivity measurement of the meltpool. The measurement equation for emissivity will be discussed in the Section 2.1.

2.1 Measurement equation

The emissivity measurement equation for each single pixel of the focal plane array (FPA) is as follows:

$$\varepsilon_\lambda(\lambda_o, T_s) = 1 - C_\rho \rho_{ref} \frac{S_{samp}^{LED\ On}(\lambda_o,T_s) - S_{samp}^{LED\ Off}(\lambda_o,T_s)}{S_{ref}^{LED\ on}(\lambda_o) - S_{ref}^{LED\ off}(\lambda_o)} \quad (1)$$

where ε_λ is the spectral normal (or 8°) spectral emissivity of the sample, λ_o is the central wavelength of the radiometer (and the light source), T_S is the sample temperature, and ρ_{ref} is the reflectance of the calibrated reference standard. The linearized signal measured by a single pixel of the imager is denoted by *S*. The signal linearization approach and its uncertainties will be discussed in Section 2.3.1. The superscripts "LED on" and "LED off" refer to signals obtained with and without LED illumination. The subscripts "samp" and "ref" refer to the objects of measure: the sample and the calibrated reflectance standard, respectively. The term C_ρ has a nominal value of one and is used as a correction factor for systematic biases. It has an associated uncertainty, which is propagated into the uncertainty of ε_λ. The correction factor C_ρ is comprised of multiple correction factors, which are associated with each source of bias, non-ideality, and uncertainty associated with the reflectometer:

$$C_\rho = C_{TU} C_{PL} C_{LED} C_{ARM} C_{AS} C_{BRDFS} C_S \quad (2)$$

where C_{TU},C_{PL}, C_{LED}, C_{ARM}, C_{AS}, C_{BRDFS},and C_S are the factors associated with throughput uniformity, port losses and high angle losses, LED reproducibility, alignment of the reference mirror, alignment of the sample, the sample bi-directional reflectance distribution function (BRDF, discussed in Section 3.1.4), and out of field scatter, respectively. Section 2.2 will describe the important design considerations of the reflectometer used for these measurements.

2.2 Baffle-less center-mount reflectometer design features

The design of the reflectometer is constrained by the size of the build chamber, the necessary gas flow provisions for laser-metal interaction, fabrication limitations, and the likelihood of damage to the reflective coating by laser reflection from the melting process. In the current case, a hemispherical reflectometer design is used instead of a complete sphere in order to address the unique considerations of the LPBF environment, while maintaining illumination performance. Use of integrating hemispheres have been applied for compact size and other optical considerations, but have not yet been applied to emissivity measurement of the LPBF to the best of our knowledge [8].

With respect to port size ratio, use of a hemisphere instead of a complete sphere allowed for a smaller laser-entrance port size and sample port size, relative to the total integration area, while fitting within the height of the build chamber of the AMMT. An equivalent height complete integrating sphere would have had about double the port area to integration area ratio. With respect to coating damage, the hemispherical design allowed the integrating surface to have greater average distance from the process, ranging from the same distance at the entrance port to double the distance at about 22° from horizontal. At normal incidence, irradiance (radiant flux received by a surface in units of W/m^2) is proportional to the square of distance, and so the hemispherical design reduces the diffuse coating exposure to intense laser reflections on average by about a factor of two. Furthermore, the hemispherical geometry allowed for double the perimeter for LED illumination, also facilitating double the intensity of illumination which must be on the same order of magnitude of the process self-emission.

A cross-sectional view of a computer aided design model of the integrating hemisphere is shown in Figure 1. As shown, optical integration is facilitated by a diffuse, barium sulfate coating and with a specular, polished aluminum base electroplated with gold, which has excellent reflectivity at 850 nm. The reflected light from the hemispherical illumination is imaged through an elongated port on the top of the hemisphere, about 8° from the vertical. The 8° offset of the port prevents retroreflection from a specular (or nearly specular) sample into the inline optics, reducing the possibility of incomplete and/or nonuniform illumination of the sample due to the detection port. Furthermore, in the case of the current application, reduced likelihood of retroreflection from the sample into the inline optics also reduces the possibility of damage to the filters or the imager from retroreflected high-power laser light. The elongated design of the laser port and sample port enables scans to be performed within an area of about 3 mm × 20 mm with coaxial imaging of the laser-metal interaction scene.

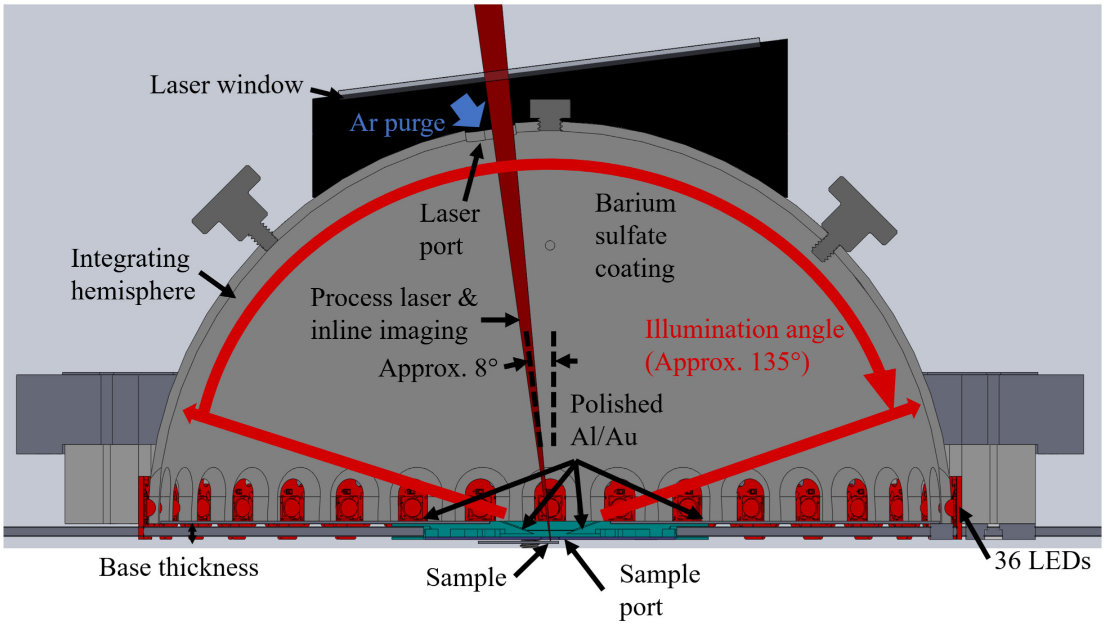


Figure 1. Cross-sectional view of a computer aided design model of the integrating hemisphere.

Another design constraint unique to the LPBF environment is that inert gas atmosphere is required for laser melting in order to reduce detrimental oxidation [9(p3)]. Previous studies have shown that directional and inert shield gas flow is essential to facilitate continuous, consistent beam delivery by removing process biproducts that can distort, scatter, and obstruct beam delivery [10,11]. As shown in Figure 1, Ar gas is pumped into the reflectometer through the laser port. The gas flow rate is typically about 30 L/min, which results in a downward flow onto the sample of about 0.5 m/s, which provides some byproduct removal. Future tests will incorporate an improved directional gas flow provision. In the current investigation, laser scanning occurred in the direction perpendicular to the image plane of Figure 1.

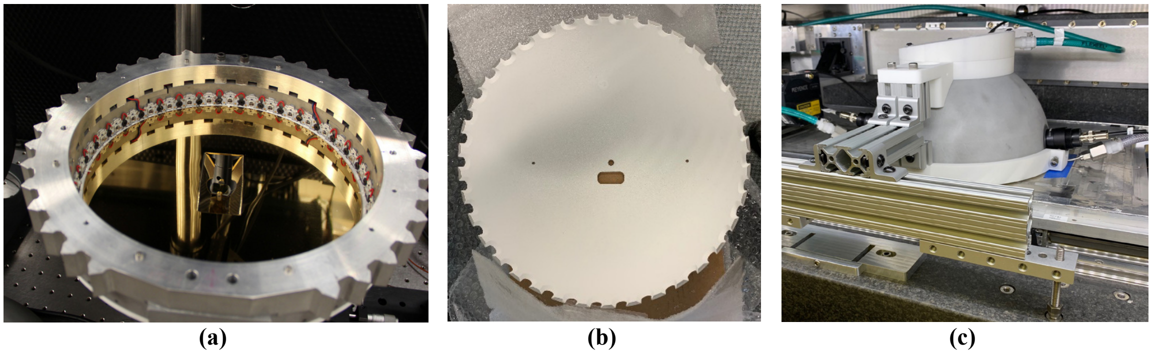


Figure 2. Images of the fabricated integrating hemisphere: (a) the specular base assembly, (b) the barium sulfate coated hemisphere interior, and (c) the assembled apparatus mounted in the build chamber.

The surfaces near the sample port are angled at about 22°, as shown in Figure 1. This creates a base thickness of about 6.7 mm. The thickness of the base of the dome acts as a baffle and results in an illumination angle of about 135°, as shown in Figure 1. The LEDs are located as close as possible to the equator of the hemisphere. At this location, the base thickness has a beneficial baffle-effect, and prevents deleterious direct illumination of the sample by the LEDs, which would lead to nonuniform illumination. The base thickness, though, causes a loss of light in the remaining 45° of the hemisphere, which does not contribute to illumination. In practice, surface features reflecting light into the directional imaging path at an angle less than 22° from the horizontal are unlikely, but possible, and the associated contribution to surface reflectance/emissivity measurement uncertainty must be evaluated. The fabricated base, hemisphere interior, and assembled integrating hemisphere within the build chamber are shown in Figure 2. Each source of error and uncertainty associated with the reflectometer-based measurement of emissivity will be described in Section 2.3.

2.3 Directional imaging signal corrections

The corrections used to condition the signal used for the measurement of spatially resolved emissivity will be described at length in a forthcoming publication, hence these items will be only briefly summarized in this section. Two primary corrections are applied to the signal: 1) the erroneous signal components due to stray light and blooming are subtracted, and 2) the images are deconvolved with a blur kernel to compensate for image distortion due to the optical blur and finite pixel size of the FPA. The corrections are applied to each image of the 30 central images of the test sequence to avoid transients due to startup and laser shutoff.

In this work, motion blur is not considered because the image is static relative to the laser motion, making the meltpool quasi-static in this reference frame. Although the meltpool undulates slightly relative to the laser motion, blur due to meltpool length changes and location variability within the reference frame within the integration time is assumed to be negligible.

2.3.1 Linearization with radiant flux

The imager signal must be linearized with respect to the object radiant flux for accurate determination of the emissivity. The imager signal is linearized by exposure to the steady, known radiance of a high temperature blackbody (HTBB). The imager is outfitted with a long working distance microscope lens body and objective lens, a band filter, and a laser cutoff filter. The measurement equation for linearization is then as follows:

$$S(T)_{cal} = C_{cal,BB} \int_{\lambda 1}^{\lambda 2} \tau_{\lambda}^{obj}(\lambda) \tau_{\lambda}^{filt}(\lambda) r_{\lambda}^{FPA}(\lambda) \frac{c_{1L}}{\lambda^5 \left[exp(c_2/(\lambda T_{rad,BB})) - 1 \right]} d\lambda \tag{3}$$

where the pixel signal $S(T)_{cal}$ is the FPA signal, $\tau_{\lambda}^{obj}(\lambda)$ is the spectral transmittance of the objective lens, $\tau_{\lambda}^{filt}(\lambda)$ is the filter spectral transmittance, and $r_{\lambda}^{FPA}(\lambda)$ is the imager spectral responsivity, c_{1L} and c_2 are the first and second radiation constants, λ is wavelength, and $T_{rad,BB}$ is the radiance temperature of the HTBB. The signal is linearized with the single-point averages of image sequences at varying HTBB temperatures. This can be done because of the relatively low noise and FPA nonuniformity, both of which are incorporated into the measurement uncertainty. Once $C_{cal,BB}$ is determined at a single point, the signal at each HTBB temperature is linearized using Equation (3) to generate a linearization function for the signal.

2.3.2 Stray light and blooming correction

Stray light is light that passes through the optical system in a manner that is not intended in the optical system. Imager blooming occurs when a pixel potential well is overfilled and the excess charge bleeds over into adjacent pixels. Both stray light and blooming tend to increase the FPA signal near high intensity regions, typically exhibiting exponential decay in erroneous signal with distance from the high intensity regions. Although the TEMPS optical system is designed using best practices to reduce stray light [12] and the imager with complementary metal–oxide–semiconductor (CMOS) is relatively impervious to blooming, these effects must be measured and compensated for high accuracy measurements. Our measurements indicate that the meltpool hotspot generates about 100 times more radiant flux than the intensity levels of interest near the melting temperature, which produces a significant potential for erroneous signal due to stray light and blooming.

Stray light and blooming are be measured simultaneously because both have similar causes and effects on the measurements. In order to quantify the combined effect of stray light and blooming, an illumination source other than the meltpool is required to eliminate the optical effects of the laser-melting process variability and byproducts. The source chosen for the application is a 6 µm diameter fiber coupled laser with divergence angle exceeding the acceptance angle of the TEMPS optics. The fiber-coupled laser is located under an aperture located at the build plane height so that the aperture is slightly overfilled. Illumination is projected onto the imager through the TEMPS optics system to replicate the meltpool hot spot. The aperture size and intensity are selected to closely mimic the area and intensity of the meltpool hotspot.

A curve fit is performed on the erroneous signal generated outside of the 100 µm aperture. The center of the hotspot is then located and the distance of the center of each pixel from the hotspot center is calculated. The erroneous signal at each pixel due to stray light and blooming is calculated from the curve fit based on the pixel distance from the hotspot. Finally, the erroneous signal is subtracted from each meltpool image of interest. The pixels in the saturated region are left at the original saturated digital level (DL). After subtraction of the signal due to stray light and blooming, a mild smoothing operation is then applied to images.

2.3.3 Image smoothing

A mild smoothing filter is applied to each test image after the stray light and blooming subtraction. Smoothing the images reduces the effects of noise on the deconvolution. The smoothing filter is based on nearest-neighbor averaging with one adjacent pixel on either side, top and bottom. We confirmed that the smoothing operation does not introduce a systematic signal bias by subtracting the original image from the smoothed image and averaging across the frame. The resulting average signal difference has been found to result in a negligible average bias of less than 1% of a digital level at each pixel.

2.3.4 Image deconvolution

Deconvolution is an image processing operation intended to reconstruct an image to its original form prior to blur induced by inevitably non-ideal optics and finite pixel size of FPAs. The multistep deconvolution approach taken here is based on that of Lane and Whitenton [13] and ISO 12233:2017 [14]. First, a knife edge measurement is taken to establish an edge spread function (ESF). The ESF is then transformed into a point spread function (PSF), which can be thought of as a "deblurring kernel." Finally, the images are deconvolved. The details of each element of the operation are discussed.

2.3.4.1 Knife-edge measurement

Use of an ESF measurement is a practical approach to the determination of the PSF, because use of a true point source is experimentally difficult, if not impossible [14]. Use of an ESF to determine the PSF reasonably assumes that the response of the optical system is rotationally symmetric.

To establish the ESF, a thin, opaque, and straight edge is placed to partially cover the aperture of the thermal integrating sphere source with 850 nm center wavelength (TISS850) set at a radiance temperature of 1600 °C at the build plane. The TISS850 is transfer source developed in the AMMT that is composed of LEDs that are thermally stabilized (by ethanol heat pipes cooled by a thermoelectric cooler and fan) to illuminate a polytetrafluoroethylene (PTFE) integrating sphere that generates uniform illumination at the waveband of interest—brightness that is calibrated against the HTBB thermal source. The edge is placed a few degrees from vertical so that it is not perfectly aligned with the pixel array [13,14]. Sampling lines are then taken perpendicular to a curve fit of the maximum image gradient to mark the edge between the high and low DLs. The image data are taken with a resolution of 6.0 µm per pixel through the TEMPS optics and the imager is set to a shutter speed of 98.3 µs, which is followed by determination of the ESF.

2.3.4.2 Edge spread function determination

The four sampling lines across the knife-edge image are found to be negligibly different, and so one of them is taken as representative of the image profile. The representative profile is then normalized to have a peak intensity of unity and is supersampled with 4 samples per pixel along the length of the array, per ISO 12233:2017 [14]. The data is then centered on zero based on the maximum gradient of the profile. In sampling with one, two, and three component error function fits, it is found that the most appropriate fit is with the two-component error function shown in Equation (4), with $R^2 = 0.998$.

$$\mathrm{ESF} = a_1 \operatorname{erf}\frac{x}{b_1} + a_2 \operatorname{erf}\frac{x}{b_2} + 0.5 \tag{4}$$

where x is distance in pixels, $b_1 = 3.184$ [pixels], $a_2 = 69.04$, and $b_2 = 3.184$ [pixels].

2.3.4.3 Point spread function determination

The PSF is determined from the ESF via differentiation followed by an Abel Transform as per Lane and Whitenton [13]. The corresponding PSF is shown in Equation (5).

$$\mathrm{PSF} = \frac{2}{\pi}\frac{a_1}{b_1^2}\exp\left(-\frac{r^2}{b_1^2}\right) + \frac{2}{\pi}\frac{a_2}{b_2^2}\exp\left(-\frac{r^2}{b_2^2}\right) \tag{5}$$

where r is the radial distance in pixels, and a_1, b_1, a_2, and b_2 are the values determined from Equation (4). The value of the PSF array is calculated at each super-sampled pixel location using Equation (4), and then averaged across the 4 × 4 super-sampled pixel to determine the PSF value at each pixel. The resulting PSF array is 25 pixels × 25 pixels in order to cover four orders of magnitude of intensity from the PSF center to the perimeter. The volume under the PSF array is finally normalized to unity.

2.3.4.4 Deconvolution

The final step is the deconvolution operation. Each image of the central 30 images of the test sequence is individually deconvolved to allow for later uncertainty analysis of the deconvolved image transient variability. Each image is first corrected for stray light and blooming, then smoothed as discussed in the preceding sections. The PSF array is then used to deconvolve each image with the MATLAB[2] 'deconvlucy' function, which is based on the iterative Richardson-Lucy method [15,16].

It is observed that the image is "sharpened" after deconvolution. The sharpening causes an apparent narrowing of the meltpool and steepening of the signal gradients on the left and right side, as well as at the nose. Toward the end of the tail, the local signal gradient becomes shallower and longer along the length of the meltpool, but steeper in the transverse direction across the tail.

2.3.5 Summary of the effects of signal corrections

The effect of each image correction operation is shown by a central cross-sectional profile along the length of the meltpool in Figure 3. Beginning with a linearized signal, the stray light and blooming subtraction operation reduces the signal at the nose of the meltpool because of its proximity to the hotspot, but has a very small effect on the signal at the tail because of its larger distance from the hotpot. The smoothing operation has very little effect on the signal levels but reduces the apparent pixel to pixel noise.

Deconvolution tends to have the most notable effect on the signal levels, both at the nose and tail of the meltpool. Deconvolution sharpens the profile, reducing the signal at areas of high curvature (large values of the second derivative), and steepening the profile at areas of high (two-dimensional) gradients. This can be observed at the nose and tail of the meltpool at pixel 10 and 48, respectively. The most significant variation in the central cross-sectional profile due to deconvolution occurs in the tail, where a "solidification plateau" appears. At this solidification plateau from pixel 59 to 63, the signal is flat, indicating an apparently isothermal solidification region as the laser heat source moves. This isothermal region is expected in the solidification process of high purity metals (99.998% Ni used here), because the fixed-point freezing temperature at which the phase change occurs is maintained as the material dissipates the latent heat of fusion at the solidification temperature. The solidification region is made significantly more apparent in the signal by deconvolution. The signal uncertainty incurred by each signal correction operation will be discussed in Section 3.

[2] Certain commercial entities, equipment, or materials may be identified in this document to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

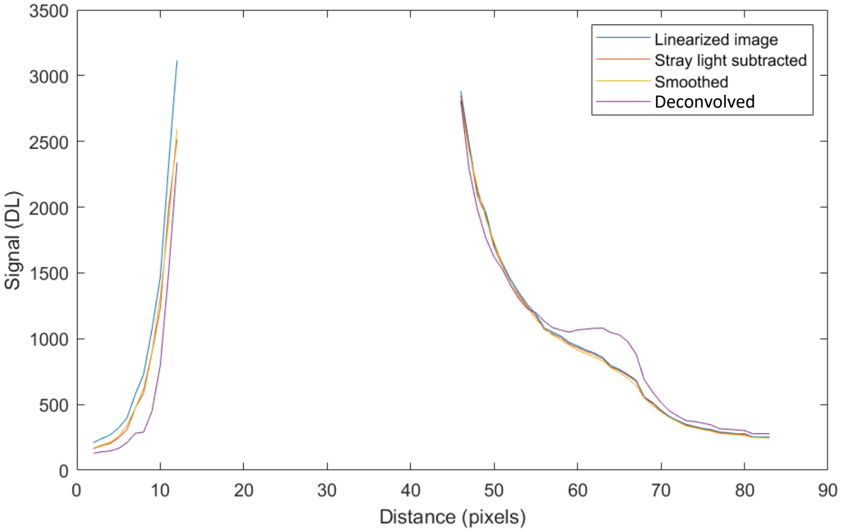


Figure 3. Central cross-sectional profile along the length of the meltpool showing the effect of each image correction. Meltpool nose is on the left and tail is on the right.

3. UNCERTAINTY

Three categories of sources of emissivity uncertainty have been identified. The first includes the nonideality, misalignment, and uncertainty in reflective character of the reference mirror and the sample. The second includes nonuniformity and incomplete hemispherical illumination, which pertains to the nonideality of the reflectometer. The third and final category of uncertainty is the nonideality of the directional imaging system, including the inline optics, imager, and image corrections.

3.1 Sample and reference uncertainties

3.1.1 Reference mirror alignment

As was discussed previously, both specular and diffuse reference samples are used in emissivity measurement and evaluation of uncertainties. The angular alignment and distance from the reflective surface to the reflectometer sample port (gap) slightly alter illumination of the reflector, as well as the reflectometer throughput (the throughput is the ratio of the flux reaching the detector to the input flux from the source). In the case of the specular mirror, misalignment or increased gap changes the location on the integrating surface from which the sample is illuminated, while also reducing throughput due to light loss from the gap. In the case of the diffuse reflectance standard, a misalignment or increased gap size changes the magnitude and location of the solid angle from which the reflector is illuminated. An experiment was performed to measure the relative change in signal when the specular reflectance standard was moved relative to the floor of the reflectometer (located at 219.3 mm from the laser window), as shown in Figure 4.

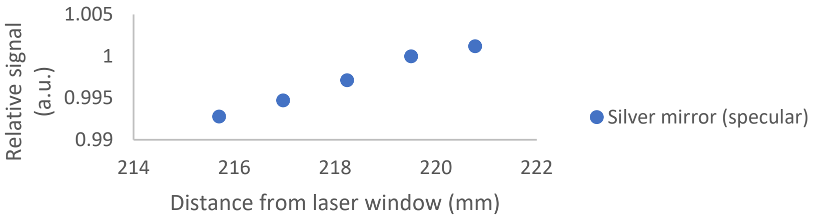


Figure 4. The relative signal measured as a function of the specular reflectance standard location relative to the laser window, with the floor of the reflectometer located at 219.3 mm

The reference mirrors are aligned with a laser displacement meter, which has an uncertainty of 1 µm (k = 1), and the reflectometer is mounted on kinematic mounts for accurate repositioning. The rotational misalignment is assumed to be negligible. It is conservatively estimated that the reflectance standard is located within ±1 mm, for which the relative range of the signal is 0.6 % (or equivalently ±0.3 % of the midpoint within the relative range). The signal has a uniform probability of being within that range, so dividing by $\sqrt{12}$, the signal uncertainty is 0.2 % [17]. Therefore, the nominal value of C_{ARM} is 1.0 with a relative standard uncertainty of 0.002.

3.1.2 Reference mirror reflectance

The reflectance standards used in these experiments are calibrated by NIST. The uncertainty in reflectance is 0.5 % [18]. Therefore, the nominal value of ρ_{ref} is 0.97 with a relative standard uncertainty of 0.005.

3.1.3 Sample gap and alignment

The effects of sample gap changes and misalignment are similar to those for the reference mirror—changes in this affects where the location from which illumination occurs on the integrating surface is, and thus the resulting throughput of the reflection. Currently, the best approximation of the effect of sample alignment is from measurements taken with a diffuse reflective standard sample (pressed polytetrafluoroethylene PTFE) and a sample with a diffuse-directional reflective character between perfectly specular and diffuse (brushed stainless steel). The results of gap change between the reflectance standards and the laser window (with the reflectometer floor located at 219.3 mm from the laser window), is shown in Figure 5.

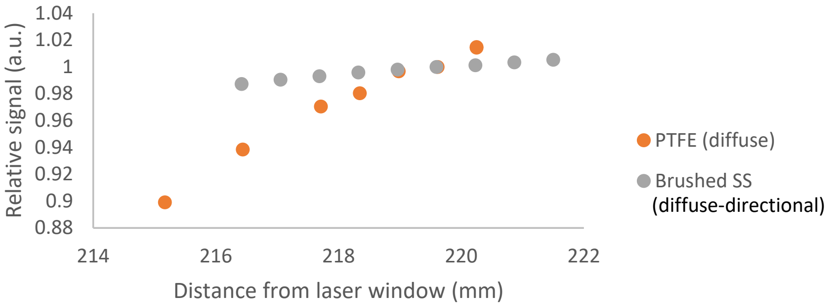


Figure 5. Relative signal measured as a function of the location of two reflectance standards relative to the laser window, with the floor of the reflectometer located at 219.3 mm

The samples of high-purity Ni for generating laser-induced meltpools are aligned with the laser focus as described above, but with a tolerance of ±20 µm (vertical misalignment is generated by surface non-flatness, general roughness, and tilt). Rotational misalignment is again assumed to be negligible. Assuming location within ±1 mm as above, and that the diffuse reflectance standard is representative of the maximum change in reflectance of a real sample (e.g., stainless steel reflectance sample, or laser-induced melt pool), the relative range of signal is about 5.0 % (or equivalently, about ±2.5 %

from midpoint of the range). Assuming a uniform distribution within this range, the relative standard uncertainty of C_{AS} becomes 1.5 %. Therefore, $C_{AS} = 1.0 \pm 0.015$.

3.1.4 Sample bidirectional reflectance distribution function

The sample BRDF is a mathematical function that describes how the hemispherical illumination of the sample is reflected and imaged by the directional imaging system. Currently, very little is known or has been measured regarding the reflective character of the laser-metal interaction scene and is an important area for future investigation. In the absence of additional information, the sample BRDF is assumed to have a similar effect to the throughput uniformity (discussed in Section 3.2.1) with no bias. Therefore, $C_{BRDFS} = 1.0 \pm 0.01$. It should be noted that under certain specific (relatively unlikely) circumstances, the sample BRDF could be a dominant uncertainty component.

3.1.5 Summary of sample and reference uncertainties

The uncertainties due to the sample and reference conditions are summarized in Table 1, where C_{ARM}, C_{AS}, C_{BRDFS}, and C_S are defined in Section 2.1. The reflectance of the reference mirror is ρ_{ref}.

Table 1: Summary of uncertainties due to sample and reference conditions

Variable	*Value*	*Units*	*Relative uncertainty (k = 1), (%)*	*Type*
C_{ARM}	1	-	0.2	A
C_{AS}	1	-	1.5	B
C_{BRDFS}	1	-	1.0	B
C_S	0.9975	-	0.5	B
ρ_{ref}	0.97	-	0.5	A

3.2 Nonuniform and incomplete hemispherical illumination

3.2.1 Throughput uniformity

Reflectometer throughput is the ratio of the flux reaching the detector to the input flux from the source. Relative throughput is measured across the integrating surface and reported in arbitrary units. Relative throughput mapping of the inside of the hemisphere quantifies the uniformity of hemispherical illumination of the reflectometer as a whole and allows estimation of the measurement uncertainty. Throughput mapping is performed by placing a photodetector at the entrance port and a gimbal-mounted mirror is located at the sample port. The mirror is aimed at a representative number of locations across the inside of the hemisphere.

As shown in Figure 6, the throughput uniformity of the surface is within ±1.0 % across from about 5° from the vertical to about 80°. Decreased radiance throughput occurs at the imaging port (5° to 15°), at the monitoring diode ports (45°), and increased high angle losses occur opposite to the port at 70° to 90°. The main features and nonidealities of the reflectometer are detected at these locations, and the uniformity around these features is taken as representative of the remainder of the reflective surface. From these results, the throughput uniformity we can assume a negligible bias in the emissivity calculation and the uncertainty is taken to be 1.0 %. Therefore, the nominal value of C_{TU} is 1.0 with a relative standard uncertainty of 0.01.

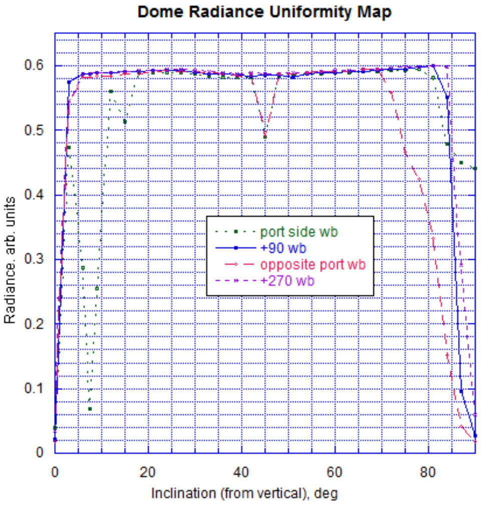


Figure 6. Relative throughput of the hemispherical integrating surface for four orthogonal cross sections. The tests are performed with the specular gold base (wb).

3.2.2 Port losses and high-angle losses

The light loss due to the laser port and base thickness on the first reflection from the sample are measured by comparing the measured signal with a specular silver standard and a diffuse gold standard. On the first reflection, a specular sample does not result in any port loss or high angle loss. In contrast, a perfectly diffuse sample results in both high angle and port losses. After accounting for the reflectivity of the samples, it is found that 2.5 % in intensity is lost for the diffuse sample as compared to the specular sample. The laser-metal interaction scene is likely better represented by a value in between the specular and diffuse cases, so the bias is assumed to be of the average of the losses for the two samples plus or minus half of the difference. Then the nominal value of C_{PL} is 1.013, with a standard uncertainty of 0.013.

The reflective character of the laser-metal interaction scene is a significant unknown and could potentially result in greater port losses and high angle losses than those considered here. Therefore, two approaches have been identified to better quantify the emissivity uncertainty component due to port losses and one approach has been identified for estimating high angle losses.

The first approach to estimating port losses is to reduce the size of the port to as small as possible, which would require a shortened laser scan distance. The relative change in signal compared with the normal port size can then be used to calculate a more representative port loss. The second approach is to add a supplementary light source via a beam splitter directed along the imaging path to add additional light from the port area, which will then be optically integrated in the reflectometer. The additional signal may then be used to better quantify port loss bias and uncertainty.

To estimate high angle losses, a small sample is moved upward through the base of the reflectometer with an adjustable cylindrical baffle surrounding it. The motion of the sample relative to the baffle and reflectometer base and the resulting change in signal intensity will be used to quantify the effect of high angle losses. The high angle losses are expected to be negligible under most circumstances, but the high-angle loss test raises concerns about scattering and absorption in the process plume and how to deconvolve those effects from high-angle loss effects, which must be addressed. Nevertheless, the combined port loss and high-angle loss estimate reported here is should be acceptable under most conditions.

3.2.3 Coating reflectance and diffuseness

Both the reflectance and the specularity of the integrating sphere coating can have an effect on the size of the potential measurement error and resulting uncertainty, with more reflective and more diffuse coatings resulting in more accurate measurements [19,20]. The barium sulfate coating used in this application has a highly diffuse reflectance with a directional-hemispherical reflectance of 0.981 at 850 nm [21]. Therefore, the coating reflectance and diffuseness is expected to have a negligible effect on the accuracy of measurements in the given application and are therefore omitted from Equation (1) and the uncertainty budget.

3.2.4 LED reproducibility

The intensity and spectrum of the illumination LEDs change as the junction temperatures increase within the semiconductor devices. According to the manufacturer's datasheet, the radiant flux output decreases by as much as 35 % when the LED case temperature changes from room temperature to 100 °C, and the peak wavelength shifts by 13 nm [22]. The absolute shift in output and wavelength are not of high importance alone, but the reproducibility of the output from the sample and the reflectance standard introduces an additional uncertainty component in the emissivity measurement. We have tested reproducibility of the LED illumination previously and estimate that it induces an uncertainty of 1.5 % in the LED correction factor C_{LED}. Stochastic non-reproducibility does not induce a bias. Hence, the nominal value of C_{LED} is 1.0 with a relative standard uncertainty of 0.015.

The uncertainty component associated with LED reproducibility can be measured with repeat tests using the reflectance standard while recording the signal received by the imager. This uncertainty component can also be reduced with combined triggering of the LEDs with triggering of the imager and laser scanning system.

3.2.5 Summary of uncertainties due to nonuniform and incomplete illumination

The uncertainties associated with nonuniform and incomplete illumination are summarized in Table 2, where C_{TU},C_{PL}, and C_{LED} are defined in Section 2.1.

Table 2: Summary of uncertainties due to nonuniform and incomplete illumination

Variable	*Value*	*Units*	*Relative uncertainty (k = 1), (%)*	*Type*
C_{TU}	1	-	1.0	A
C_{PL}	1.013	-	1.3	B
C_{LED}	1	-	1.5	B

3.3 Directional imaging

Nine sources of uncertainty in the FPA signal are considered in addition to the aforementioned measurement uncertainty components in the measurement equation (Equation (1)), as follows:

1. Out of field scatter from sample port surfaces
2. Polarization effects
3. Linearization uncertainty
4. Uncertainty due to stray light and blooming correction
5. Uncertainty due to image smoothing
6. Uncertainty due to point spread function determination
7. Uncertainty due to deconvolution operation
8. Uncertainty due to deconvolved signal variability caused by process variability and imager noise
9. FPA nonuniformity

This section describes the uncertainty of the corrections, as well as the resulting uncertainty of the measured emissivity.

3.3.1 Out of field scatter from sample port surfaces

The edges of the reflectometer sample port are a non-ideality that may cause out of field scatter that may be detected as erroneous signal by the directional imaging system. This effect is expected to be small, and in the absence of experimental data, it is assumed to induce an increase of signal bias of 0.25 % with a (Type B) relative standard uncertainty of 0.5 %. Therefore, $C_S = 0.9975$ with a relative standard uncertainty of 0.005. The port surfaces may be coated with absorptive (black) paint in the future to reduce the associated bias and uncertainty of out of field scatter.

3.3.2 Polarization effects

Polarization may significantly alter the throughput uniformity of integrating spheres and should be considered as a potential uncertainty component [23]. The integrating hemisphere uses 36 symmetrically distributed LED sources for illumination, which means that the illumination is very unlikely to have a preferred polarization illumination of the sample or reference mirror. The samples used for meltpool generation are randomly sanded, making the un-melted material unlikely to have preferred polarization reflectivity. The Ag reference mirror has very low polarization at 850 nm and at the 8° from normal reflection angle, making reference mirror polarization bias unlikely. Finally, nearly all practically useful data are obtained on molten or re-solidified sample surfaces, which have potential for polarizing effects, but this has not yet been measured. In the future, a series of tests will be performed with varying scan direction and rotation of the randomly sanded sample to quantify polarization effects. Currently, measurement bias and uncertainty of emissivity due to polarization is assumed to be negligible.

3.3.3 Imager signal linearization uncertainty

The primary instrument used for the emissivity measurement uses a CMOS FPA. The imager has a 1024 pixel × 1024 pixel, 12-bit dynamic range FPA. In this work, all data are obtained with a shutter speed of 98.3 µs. The transient pixel noise and nonuniformity across the FPA are sources of signal variance and therefore contribute to the measurement uncertainty. Furthermore, the nonlinearity of the signal with the incident flux requires correction by calibration to a known-flux source, which introduces an additional uncertainty component for the signal. Each of these are discussed in this section.

In order to evaluate the uncertainty due to transient pixel noise, the HTBB is used as a stable source. The imager is outfitted with an long working distance microscope lens body and an objective lens focused at the HTBB aperture to generate uniform and steady irradiance of the FPA. The majority of the signal variation is therefore due to electronic noise. The nonuniformity of the blackbody irradiation is considered to be negligible in this evaluation. The GainCal function in the imager software is used for flat-fielding (or, nonuniformity correction) to reduce the natural optical vignetting and improve the pixel uniformity across the FPA.

3.3.3.1 Pixel noise and FPA nonuniformity

Samples of 100 images are taken with varying HTBB set-point temperatures. Then the standard error (SE) of the mean DL of each pixel is determined from the 100-image sample. Use of the dynamic range is limited to between 300 DL and 3800 DL, which results in a transient pixel noise relative standard uncertainty component of the signal of 0.3 %.

Images from the linearization against the HTBB are used to measure the FPA nonuniformity. The pixel average is taken from the 100-image sample, resulting in a calibration image with the average DL of each pixel. The standard deviation across the pixel array is then determined to evaluate the nonuniformity across the FPA. The resulting frame standard deviation of the pixel average is expressed as a percentage of the DL. Use of the dynamic range is limited to between 300 DL and 3800 DL, which results in an FPA nonuniformity relative standard uncertainty component of the signal of 1.0 %. The combined transient pixel noise and FPA nonuniformity relative standard uncertainty is then 1.05 %.

3.3.3.2 Spectral variable uncertainties

In order to solve Equation (3), the lens spectral transmission $\tau_\lambda^{obj}(\lambda)$, filter spectral transmission ($\tau_\lambda^{filt}(\lambda)$), FPA spectral responsivity (r_λ^{FPA}), and HTBB temperature (T_{BB}) must be measured. Inspection of Equation (3) shows that the absolute biases of the spectral quantities will simply change the value of $C_{cal,BB}$, which does not affect its uncertainty. Therefore, deviation in the magnitude of the spectral values across the waveband of interest are used to evaluate the uncertainty of $C_{cal,BB}$.

The measured spectral responsivity of the FPA is measured with an instrument with a conservatively estimated relative standard uncertainty of 2.0 %. The worst case of the value varying from its minimum value of $0.98r_\lambda^{FPA}(\lambda)$ to $1.02r_\lambda^{FPA}(\lambda)$ from 830 nm to 870 nm is then used to evaluate the change in $C_{cal,BB}$. Because of the wide and uneven spectral spacing of the data points, interpolation of the values is used. All values of the spectral quantities used in Equation (3) are evaluated at the same uniformly gridded values of λ, and their product is integrated using trapezoidal summation.

The measured spectral transmission of the combined 850 nm ± 20 nm bandpass filter and the 1000 nm laser cutoff filter is also measured, and the spectrometer used to measure the transmission is known from previous evaluations to have a relative standard uncertainty of 0.5 %. The worst cases of the value varying from its minimum value of $0.995\tau_\lambda^{filt}(\lambda)$ to $1.005\tau_\lambda^{filt}(\lambda)$ from 830 nm to 870 nm is used to evaluate the change in $C_{cal,BB}$. Finally, the transmission of the camera lens assembly is conservatively assumed to have a relative standard uncertainty of 2.0 % across the waveband of interest.

3.3.3.3 Wavelength and radiance uncertainties

The spectrometer used to measure the transmission of the filters has a spectral resolution of 0.5 nm, which we use as the uncertainty in λ. The uncertainty of the blackbody radiance ($L_\lambda^{BB}(\lambda)$), defined in Equation (6), has been determined to be 0.6 % and which will be described in detail in a forthcoming publication.

$$L_\lambda^{BB}(\lambda) = \frac{c_{1L}}{\lambda^5\left[exp(\,c_2/(\lambda T_{rad,BB})) - 1\right]} \tag{6}$$

3.3.3.4 Combined uncertainty due to linearization

The combined uncertainty of the calibration constant is calculated by first evaluating Equation (3) with each variable in its worst case, or maximum uncertainty. This establishes the sensitivity of the calibration constant to each uncertainty component. Then, the sum-square of all the uncertainty components is taken to evaluate the combined uncertainty of the calibration constant. These results are shown in Table 3.

Table 3. Combined signal and linearity uncertainty components of the calibration constant $C_{cal,BB}$.

Variable	Uncertainty (k = 1)	Type	Change from nominal value	Change in $C_{cal,BB}$ (%)
$\tau_\lambda^{filt}(\lambda)$	0.5 %	B	Increasing with λ	-0.01
			Decreasing with λ	0.01
$\tau_\lambda^{CF1}(\lambda)$	2.0 %	B	Increasing with λ	-0.05
			Decreasing with λ	0.05
$r_\lambda^{FPA}(\lambda)$	2.0 %	B	Increasing with λ	-0.05
			Decreasing with λ	0.05
λ	0.5 nm	A	Absolute increase	-0.48
			Absolute decrease	0.48
$L_\lambda^{BB}(\lambda)$	0.6 %	A/B	Absolute increase	-0.6

			Absolute decrease	0.6
$S(T)_{cal}$	1.05 %	A	Absolute increase	1.00
			Absolute decrease	-1.00
			Combined uncertainty of $C_{cal,BB}$	**1.25**

3.3.4 Stray light and blooming

The range-normalized root mean square error (RMSE) of the curve fit applied to the erroneous signal due to stray light and blooming is 0.7 %. This is used as the uncertainty of the curve fit to the stray light and blooming signal matrix, combined with the uncertainty due to noise in the knife-edge measurements. As described in Section 3.3.3.4, the uncertainty of the signal linearization operation is 1.25 % and this is taken as an additional uncertainty. The combined uncertainty of the stray light and blooming correction is then 1.4 % of the DL of each pixel of the erroneous signal matrix, which is calculated for the uncertainty of each pixel of the stray light and blooming corrected image.

3.3.5 Image smoothing

As stated previously, it is confirmed that the smoothing operation does not introduce a systematic bias. This is demonstrated by subtracting the original image from the smoothed image and averaging across the frame, resulting in a negligible bias of less than 1% of a digital level per pixel. Therefore, the uncertainty due to image smoothing is taken to be negligible.

3.3.6 Point spread function uncertainty

Uncertainty in the PSF is due to uncertainty in establishing the ESF via curve fitting of the empirical data. The linearization operation of the data incurs an initial signal uncertainty of 1.25 %. The curve fit also incurs an uncertainty of 2.3 %, which is the RMSE normalized by the range. In order to establish a PSF at the uncertainty extremes, the function is scaled in the positive and negative direction by the combined uncertainty of 2.6 %. From this, new constants a_1, b_1, a_2, and b_2 of Equation (5) are found.

The method described in Section 2.3.4.3 is then used to establish two PSF arrays at both extremes of uncertainty due to the ESF. The images are then deconvolved using the method described in Section 2.3.4.4 with each PSF. The PSF that caused the larger average pixel signal variation from the nominal image is taken as the PSF for uncertainty evaluation, although the change in signal due to either PSF is nearly symmetric. The average deconvolved image produced by the uncertainty of the PSF is then subtracted from the nominal average deconvolved image to establish an uncertainty matrix due to error in the PSF determination.

3.3.7 Deconvolution

The Richardson-Lucy algorithm is among the most robust deconvolution algorithms, but it is an iterative process designed to converge on the most likely reconstructed signal values [24,25]. Conversely, convolution is a destructive forward process that can be done with very little error. Therefore, in order to estimate the uncertainty of the deconvolution, the deconvolved image is convolved and subtracted from the unconvolved image and the magnitude of th signal discrepancy is taken as a conservative estimate of signal uncertainty due to the deconvolution operation at each pixel.

Deconvolution error also occurs due to the saturated region of the image, in which the signal values are no longer proportional to the local radiant flux. In order to reduce the error associated with deconvolution of false signal values, a border of five pixels around the saturated region are discarded. The value of five pixels is chosen because that is the radius within which about 97 % of the PSF volume is contained, and therefore false signal values should have a negligible effect outside of that radius. In future work, signal values may be extrapolated to five pixels within the saturated region to reduce the number of discarded pixels. Or, a further improvement can be implemented by measuring the meltpool at varying shutter speeds to increase the useful dynamic range of measurement, eliminating any saturated signal values.

3.3.8 FPA nonuniformity

As described in Section 3.3.3.1, the FPA nonuniformity is measured after a digital nonuniformity correction and exposure to a uniform source. The pixel average is taken from the 100-image sample, resulting in a calibration image with the average DL of each pixel. The standard deviation across the pixel array is then taken as an evaluation of the nonuniformity across the FPA. The resulting frame standard deviation of the pixel average is expressed as a percentage of the DL. Use of the dynamic range is limited to between 300 DL and 3800 DL, which results in an FPA nonuniformity relative standard uncertainty component of signal of 1.0 %.

3.3.9 Process variability and signal noise

Each image of the central 30 images of the test is deconvolved, and an average of each pixel is taken along with the standard error of the mean of each pixel. The standard error of the mean of each deconvolved pixel is then used as an uncertainty component of the signal due to the physical process variability combined with electronic image noise.

3.3.10 Combined signal uncertainty

The measurement approach requires recording of image data in four steps. The meltpool is recorded with the inline imaging system with the LEDs off, then repeated with the LEDs on. Similarly, the reference mirror is imaged with the LEDs off, then repeated with the LEDs on. With meltpool tests, the frame rate and length of the scan produced image sets containing 80 images. The meltpool requires some "development length" at the initiation of laser melting, and similarly has a cooldown period once the laser power is turned off. Because of these considerations, the first 25 images and last 25 images are not used, and the central 30 images recorded during steady melting are used.

As described in the preceding sections, six sources of uncertainty components in the signal have been identified and quantified. The first uncertainty component is due to the signal linearization operation—it should be noted that this uncertainty is estimated based on a 1.25 % uncertainty in signal values after deconvolution, which inherently assumes that the linearization uncertainty transforms linearly through the deconvolution operation. The second component is incurred by the uncertainty of the stray light and blooming correction operation. The third component is due to the uncertainty in the calculation of the PSF for the deconvolution. The fourth component is based on the discrepancy between the re-convolved deconvolved image compared with the image prior to deconvolution. The fifth component is due to the basic FPA nonuniformity across the field of view. The sixth and final component is the combined effect of signal fluctuation

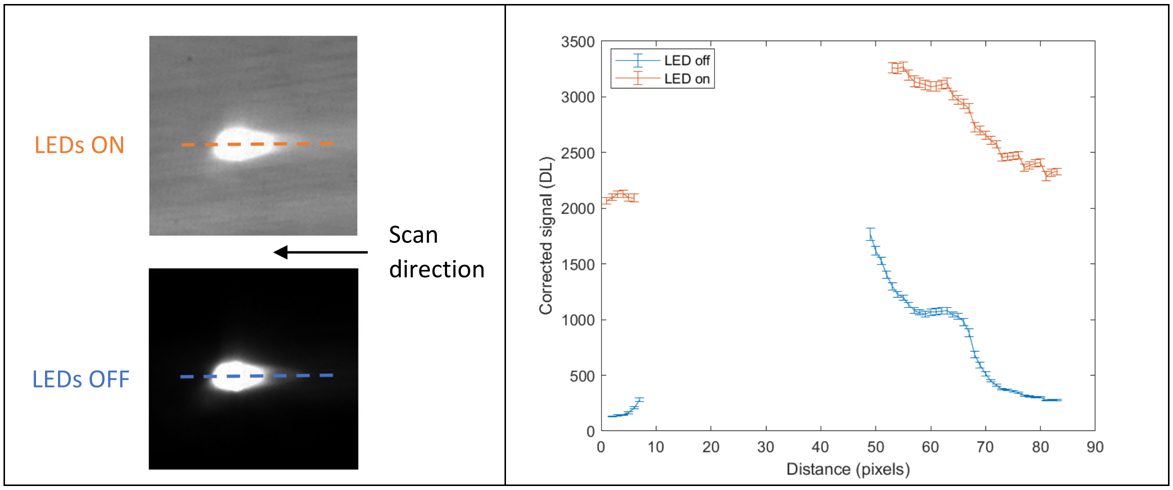


Figure 7. Image data of the meltpool generated with 99.998% Ni with the LEDs on (top left), the LEDs off (lower left), and the associated plot of the corrected signal cross-sections and uncertainties at the location of the dashed lines in the images on the left (right). The data parameters included a laser power of 250 W, scan speed of 1000 mm/s, spot size of 65 µm, image rate of 10000 Hz, and 98.3 µs shutter speed.

due to process variability combined with the electronic noise. Each of the six uncertainty components are calculated independently at each pixel to determine the local uncertainty of the signal of each pixel. The root sum-squares of the six components at each pixel is then evaluated to determine the resulting combined signal uncertainty.

Image data of the reference mirror (with the LEDs on and off) do not require stray light and blooming correction or deconvolution. Therefore, the combined uncertainty of the reference mirror images only has three components: linearization uncertainty, FPA nonuniformity, and pixel noise.

Figure 7 shows meltpool images with the LEDs off, then on, as well as central cross-sectional profiles of the corrected signal and associated uncertainty. The test parameters included a laser power of 250 W, scan speed of 1000 mm/s, spot size of 65 µm, image rate of 10000 Hz, and 98.3 µs shutter speed. The meltpool is generated with 99.998% Ni. As described previously, the signal uncertainty varies with each pixel, but typical uncertainty values are in the approximately 2 %. The solidification plateau is evident from pixel 58 to pixel 64 with the LEDs on and off in Figure 7.

4. EMISSIVITY AND UNCERTAINTY OF A HIGH-PURITY NICKEL MELTPOOL

As stated previously, the test parameters included a laser power of 250 W, scan speed of 1000 mm/s, spot size of 65 µm, image rate of 10000 Hz, and 98.3 µs shutter speed, with the meltpool generated in 99.998% Ni. The resulting emissivity values are shown in Figure 8a and the associated relative standard uncertainty is shown in Figure 8b. Figure 8c shows a central cross-sectional profile of emissivity and uncertainty along the meltpool with nose on the left and tail on the right.

Starting with Figure 8a, it can be observed that the highest emissivity values of more than 0.42 occur near the hotspot. This is likely caused by the depression in the liquid metal generated by a vapor jet emanating from the laser-metal interaction area and the resulting recoil pressure on the liquid surface. The resulting depression in the molten metal becomes a trap for illumination light by multiple reflections, and therefore decreases the local reflectivity, increasing the local emissivity. Intermediate emissivity values of about 0.4 occur in front and to the left and right of the hotspot area in Figure 8. These areas have not been melted by the laser heating, and so the original surface finish of 320 grit sandpaper grinding is maintained. The lowest values of emissivity, below 0.38, occur at the tail of the meltpool where the metal is liquid or recently solidified. These areas that have transitioned to liquid (and/or back to solid) generate more reflective surfaces because surface grinding marks have been eliminated, resulting in lower emissivity. The highest relative standard uncertainty of emissivity also occurs in the tail, which is due in large part to the uncertainty associated with deconvolution in this area of high signal gradients transverse to the meltpool.

The location of the solidification plateau is evident in Figure 7 from pixel 58 to pixel 64, which corresponds to a distance of 350 µm to 385 µm in Figure 8c. The measured value of near-normal spectral emissivity near the melting temperature (1455 °C) of 99.998% Ni is 0.36 with standard relative uncertainty of about 7 %. Published data on similar material at 1491 °C resulted in normal spectral emissivity of about 0.36 at 850 nm [26]. Therefore, under the conditions of comparison, the emissivity measurement approach developed here agrees with published values.

5. CONCLUSIONS AND FUTURE WORK

A unique high temperature reflectometer has been successfully implemented and characterized. The developed reflectometer is, to our best knowledge, the first to use an integrating hemispherical illumination setup, which has proven to be a practical approach. The high-intensity hemispherical illumination (850 nm band) used in this study produces an apparent radiance temperature of the target (with emissivity of 0.5) equal to 1705 °C, which allows for reflectometry of high temperature targets that produce significant self-emission.

This high temperature reflectometer at the NIST AMMT laboratory, with support of the TEMPS optics system and imager, facilitates the measurement of the local spectral directional emissivity, which further facilitates measurement of local surface temperatures of meltpools generated by the laser powder bed fusion process. No performance degradation of the reflectometer is observed due to damage by the high-intensity reflected laser light from the laser-melting process or from contamination by laser-melting process byproducts. Spatial distributions of the emissivity and emissivity uncertainty, of

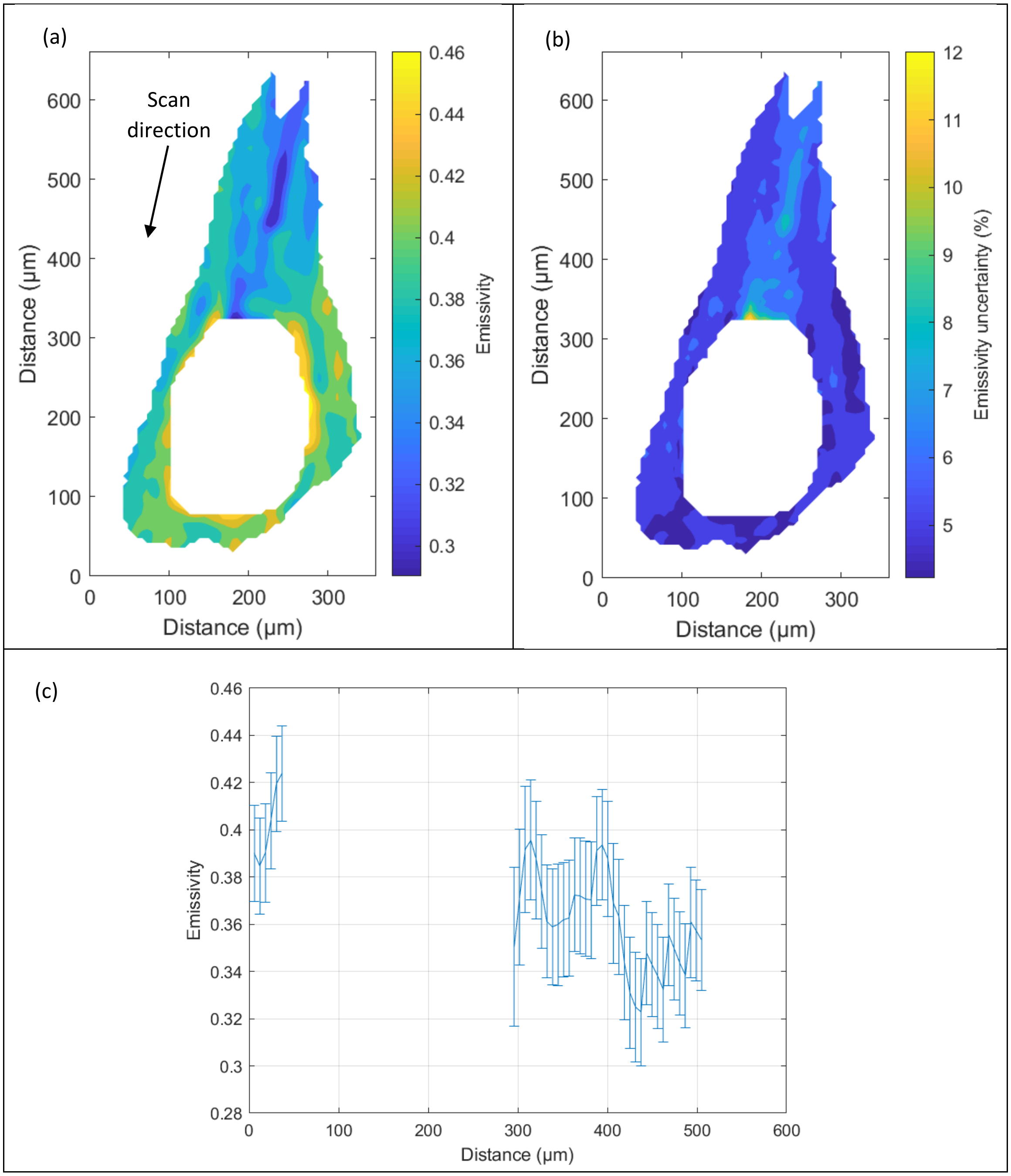


Figure 8. (a) Emissivity map of a meltpool, (b) relative standard uncertainty of the emissivity map, and (c) the central cross-sectional profile of the emissivity and the uncertainty along the meltpool with a nose on the left and a tail on the right. The data parameters included a laser power of 250 W, a scan speed of 1000 mm/s, a spot size of 65 µm, an image rate of 10000 Hz, and a 98.3 µs shutter speed with a 99.998% Ni plate.

a laser-induced meltpool of high-purity nickel, are presented. Some measurement uncertainties were estimated (Type B) in this work, but approaches are identified to better quantify the uncertainty values in the future. The measured emissivity values are found to be in good agreement with literature values at the melting point of high-purity nickel.

Due to the highly non-uniform image scene of a laser-induced meltpool, this investigation focused in part on signal processing, including signal linearization and deconvolution, and the uncertainty induced by those processing operations. Image deconvolution is found to significantly accentuate the solidification plateau of the meltpool due to the high local signal gradients, although the effect generally cancels out in the emissivity measurement equation due to an approximately equivalent effect of deconvolution with images taken with and without hemispherical illumination. Nevertheless, the current results indicate that signal processing, including deconvolution, has important ramifications for the measurement of thermodynamic surface temperatures. Approaches for estimating uncertainties associated with stray light and blooming correction, as well as deconvolution have been identified and implemented in this work.

Several hardware improvements may be implemented in the future to improve the utility of the emissivity measurement approach described here. The current reflectometer configuration uses continuous illumination by high intensity LEDs at 850 nm. The LEDs are modularized and can be changed if necessary, for measurement in other spectral bands. LEDs with shorter wavelengths will enable measurement of equivalent radiance temperatures of the probing light up to 2650 °C (for a 405 nm band illumination with a target emissivity of 0.5). Future hardware improvements also include the use of pulsed narrow band sources to increase the temperature range further, use of a thin bottom to increase the angular range of the illumination, adding the ability to translate the reflectometer, as well improvement of the shield gas arrangement to be able to work with powders. Future research also includes more detailed quantification of the port loss effects and other sources of uncertainty, especially if lower levels of uncertainty are needed. Finally, quantification of the measurement bias and uncertainty induced by laser-melting byproduct effects, such as scatter and diffraction of probing and/or reflected light by the metal vapor and condensate, will be a topic of future research.

ACKNOWLEDGEMENTS

The authors wish to thank Vladimir Khromchenko of NIST Physical Measurement Laboratory (PML) and Daniel Cardenas-Garcia of the National Metrology Center of Mexico for their work in characterizing the reflectometer. The authors also wish to thank Steve Grantham of NIST PML for his work on design and implementation of the reflectometer and LED power control. The authors further wish to thank Ivan Zhirnov formerly of NIST PML for his work in evaluation of sample/mirror alignment effects and taking of experimental data. The authors finally wish to than Brandon Lane of NIST Engineering Laboratory for his technical review of this content.

REFERENCES

[1] Herzog, D., Seyda, V., Wycisk, E. and Emmelmann, C., "Additive manufacturing of metals," Acta Mater. 117, 371–392 (2016).

[2] Mitchell, J. A., Ivanoff, T. A., Dagel, D., Madison, J. D. and Jared, B., "Linking pyrometry to porosity in additively manufactured metals," Addit. Manuf. 31, 100946 (2020).

[3] Hanssen, L. M., Mekhontsev, S. N. and Khromchenko, V. B., "Infrared spectral emissivity characterization facility at NIST," Thermosense XXVI 5405, 1–12, International Society for Optics and Photonics (2004).

[4] Lane, B., Moylan, S., Whitenton, E. P. and Ma, L., "Thermographic measurements of the commercial laser powder bed fusion process at NIST," Rapid Prototyp. J. (2016).

[5] Reijonen, J., Revuelta, A., Riipinen, T., Ruusuvuori, K. and Puukko, P., "On the effect of shielding gas flow on porosity and melt pool geometry in laser powder bed fusion additive manufacturing," Addit. Manuf. 32, 101030 (2020).

[6] Lane, B., Mekhontsev, S., Grantham, S., Vlasea, M., Whiting, J., Yeung, H., Fox, J., Zarobila, C., Neira, J. and McGlauflin, M., "Design, developments, and results from the NIST additive manufacturing metrology testbed (AMMT)," Solid Free. Fabr. Symp., 1145–1160 (2016).

[7] Yeung, H., Neira, J., Lane, B., Fox, J. and Lopez, F., "Laser path planning and power control strategies for powder bed fusion systems," Solid Free. Fabr. Symp., 113–127 (2016).

[8] Shiraiwa, H. and Sano, H., “Reference light source device used for calibration of spectral luminance meter and calibration method using same,” US10330530B2 (2019).

[9] Ahn, J., He, E., Chen, L., Dear, J. and Davies, C., “The effect of Ar and He shielding gas on fibre laser weld shape and microstructure in AA 2024-T3,” J. Manuf. Process. 29, 62–73 (2017).

[10] Malekipour, E. and El-Mounayri, H., “Common defects and contributing parameters in powder bed fusion AM process and their classification for online monitoring and control: a review,” Int. J. Adv. Manuf. Technol. 95(1–4), 527–550 (2018).

[11] Ladewig, A., Schlick, G., Fisser, M., Schulze, V. and Glatzel, U., “Influence of the shielding gas flow on the removal of process by-products in the selective laser melting process,” Addit. Manuf. 10, 1–9 (2016).

[12] Grantham, S., Lane, B., Neira, J., Mekhontsev, S., Vlasea, M. and Hanssen, L., “Optical design and initial results from NIST’s AMMT/TEMPS facility,” Laser 3D Manuf. III 9738, 97380S, International Society for Optics and Photonics (2016).

[13] Lane, B. and Whitenton, E., “Calibration and measurement procedures for a high magnification thermal camera,” Rep. No NISTIR8098, Natl. Inst. Stand. Technol. (2015).

[14] “ISO 12233:2017 Photography — Electronic still picture imaging — Resolution and spatial frequency responses.”

[15] Richardson, W. H., “Bayesian-Based Iterative Method of Image Restoration,” JOSA 62(1), 55–59 (1972).

[16] Lucy, L. B., “An iterative technique for the rectification of observed distributions,” Astron. J. 79, 745 (1974).

[17] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML., “Guide to the Expression of Uncertainty in Measurement,” JCGM 100:2008, (2008).

[18] Barnes, P. Y., Parr, A. C. and Early, E. A., “NIST Measurement Services: Spectral Reflectance,” NIST Spec. Pub. 250-48, Natl. Inst. Stand. Technol. (1998).

[19] Hanssen, L. M., “Effects of non-Lambertian surfaces on integrating sphere measurements,” Appl. Opt. 35(19), 3597–3606 (1996).

[20] Hanssen, L. M. and Snail, K. A., [Integrating spheres for mid-and near-infrared reflection spectroscopy], John Wiley & Sons, Ltd New York (2002).

[21] “Avian-B White Reflectance Coating.”, Avian Technol.

[22] LED Engin., “850nm Infrared LED Emitter LZ4-00R408” (2018).

[23] Hanssen, L., “Integrating-sphere system and method for absolute measurement of transmittance, reflectance, and absorptance of specular samples,” Appl. Opt. 40(19), 3196–3204 (2001).

[24] Fish, D. A., Brinicombe, A. M., Pike, E. R. and Walker, J. G., “Blind deconvolution by means of the Richardson–Lucy algorithm,” JOSA A 12(1), 58–65 (1995).

[25] Gladysz, S. and Gallé, R. B., “Comparison of image restoration algorithms in the context of horizontal-path imaging,” Infrared Imaging Syst. Des. Anal. Model. Test. XXIII 8355, 83550X, International Society for Optics and Photonics (2012).

[26] Krishnan, S., Yugawa, K. J. and Nordine, P. C., “Optical properties of liquid nickel and iron,” Phys. Rev. B 55(13), 8201–8206 (1997).

Calibration of free-space and fiber-coupled single-photon detectors

Thomas Gerrits[1], Alan Migdall[2,3], Joshua C. Bienfang[2], John Lehman[1], Sae Woo Nam[1], Oliver Slattery[3], Jolene Splett[1], Igor Vayshenker[1], Jack Wang[1]
[1]National Institute of Standards and Technology, Boulder, CO, 80305 (USA)
[2]Joint Quantum Institute, University of Maryland, [3]National Institute of Standards and Technology, Gaithersburg, MD, 20899 (USA))
Author e-mail address: gerrits@nist.gov

Abstract: We present our measurements of the detection efficiency of free-space and fiber-coupled single-photon detectors at wavelengths near 851 nm and 1533.6 nm. We investigate the spatial uniformity of one free-space-coupled silicon single-photon avalanche diode (SPAD) and present a comparison between fusion-spliced and connectorized fiber-coupled single-photon detectors. We find that our expanded relative uncertainty for a single measurement of the detection efficiency is as low as 0.7 % for fiber-coupled measurements at 1533.6 nm and as high as 1.8 % for our free-space characterization at 851.8 nm. © 2020 The Author(s)

Future Optical Quantum Networks will need components based on single-photon quantum technologies and those components will require characterization. We start with single-photon detectors, which in turn can be used to characterize other quantum network components such as single-photon sources, fiber losses, network switches, etc.

We measure detection efficiency using a calibrated attenuation stage and a calibrated optical power meter as shown in Fig. 1(a) [1]. Laser power is first roughly set using a variable fiber attenuator (VFAinput) and then sent to a splitter/attenuator unit, which has a highly attenuated output and high-light-level monitor. The ratio of the output to the monitor (Rout/mon) is about 10^{-5} and is measured using an optical power meter (PM) and monitor optical power meter (PMmon). Both, PM and PMmon, require a nonlinearity (relative) calibration, whereas only PM requires an absolute responsivity calibration. Key to the measurements are the transmittance of the splitter/attenuator unit and the output-to-monitor ratio of the splitter/attenuator unit. Both are determined from the fiber beam splitter (FBS) splitting ratio and the attenuation of VFA, using the power meter and the monitor power meter. In addition, this method relies on the stability of the splitter/attenuator unit's output-to-monitor ratio, the polarization and wavelength of the light versus time, and the independence of the output-to-monitor ratio with input optical power. We verify each of these either during the measurement or by prior characterization.

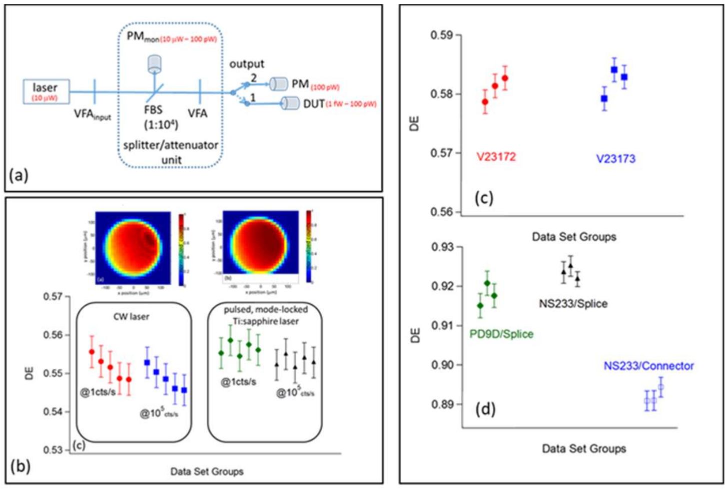


Fig. 1. (a) Schematic of the measurement setup. (b) Measured DE for the NIST8103 detector at 1 cnt/s and 10^5 cnt/s made with CW laser and a pulsed Ti:sapphire laser, as labelled. Error bars represent the extracted standard uncertainties (*k*=1) for each measurement. (c)-(d): Summary of measurements for fiber-coupled detectors (c) V23172, V23173, and PD9D at 851.8 nm and (d) NS233 at 1533.6 nm. For detector NS233, fusion splicing and FC/PC connectors were used to connect the DUT fiber to the output of the FBS as indicated. Error bars are the extracted standard uncertainties (*k*=1) of each measurement

Gerrits, Thomas; Migdall, Alan; Bienfang, Joshua; Lehman, John H.; Nam, Sae Woo; Slattery, Oliver T.; Splett, Jolene D.; Vayshenker, Igor; Wang, Chih-Ming. "Calibration of free-space and fiber-coupled single-photon detectors." Paper presented at Quantum 2.0 Conference. September 14, 2020 - September 17, 2020.

Detection efficiency results are shown (Fig. 1(b)) for a free-space-coupled silicon single-photon avalanche detector (SPAD) measured in two modes. One using a continuous wave (CW) and another using a mode-locked Ti:sapphire laser. Detection efficiency (DE) is measured at a range of detector count rates so that the DE can be determined at 1 cnt/s and 10^5 cnt/s. Setup stability and repeatability is achieved for the extracted detection efficiencies at the two rates of 1 cnt/s and 10^5 cnt/s (Fig. 1(b)).

The CW laser results show a larger variation in the extracted DE at both count rates. The spatial response of the SPAD using the CW and Ti:sapphire laser, measurements are also shown. The CW laser results show fringes in the spatial response. Thus, it will have higher sensitivity to slight spatial misalignments than for the measurements made with the Ti:sapphire laser.

Figure 1(c) shows the extracted detection efficiencies at 10^5 cnt/s of three fiber-coupled detectors, two SPADs: V23172 and V23173 and one superconducting nanowire single photon detector (SNSPD): PD9D at a wavelength of 851.8 nm. The DEs for both SPADs were determined with an FC/PC fiber connector at the output fiber of the FPC, whereas the SNSPD's DE was determined by fusion splicing the detector fiber. Good setup reproducibility is observed for all three detectors. Figure 1(d) shows the calibration results of an SNSPD optimized for 1550 nm (NS233) at a measurement wavelength of 1533.6 nm. The DE of NS233 was determined with an FC/PC fiber-to-fiber connector union and by fusion splicing the detector fiber to the output fiber of the FPC. The measured extracted DE at 10^5 cnt/s through a fusion splice is higher than that measured through an FC/PC connector, as expected. The repeatability between individual runs for both cases is comparable to the repeatability achieved for the 851.8 nm fiber-coupled measurement.

Table 1. Summary of results for all measured single photon detectors. Quoted are the photon delivery method (fiber-coupled, free-space, Ti:sapphire laser, CW laser, fusion spliced, connectorized, direct fiber connection), mean DEs at 10^5 cnt/s, the 95 % coverage intervals and the relative expanded uncertainties (k=2).

detector	CW	Ti:sapphire	Free-space	Direct fiber	connectorized	Fusion-spliced	DE at 10^5 cnt/s	95 % cov. int.	rel. exp. unc. (%)
NIST8103		x	x				0.5532	[0.5449, 0.5615]	1.5
NIST8103	x		x				0.5490	[0.5397, 0.5587]	1.8
V23172	x			x			0.5811	[0.5708, 0.5911]	1.8
V23173	x			x			0.5821	[0.5735 0.5911]	1.6
PD9D	x					x	0.9178	[0.9066, 0.9292]	1.1
NS233	x				x		0.8921	[0.8859, 0.8996]	0.73
NS233	x					x	0.9234	[0.9171, 0.9298]	0.70

Table 1 summarizes the results of this work for all detectors at a count rate of 10^5 cnt/s. The DE and 95 % coverage intervals were calculated with the NIST consensus builder [2] and linear opinion pooling for the individual measurement outcomes for each detector. Relative expanded uncertainties as low as 0.70 % are achieved in the case of a fiber-coupled SNSPD at 1533.6 nm. Whereas for the free-space measurements at 851.8 nm, the relative expanded uncertainty is 1.8 % with a CW laser. The main source of uncertainty for the free-space measurements is the uncertainty in the detector response due to laser-beam-detector alignment. For all-fiber-coupled detectors this uncertainty is not relevant but is replaced with a connector and fiber-end reflection-loss uncertainty. In this study, we were not able to compare several FC/PC connectors to establish an uncertainty associated with different commercially available fiber connectors. However, we believe that for many different FC/PC connectors the loss uncertainty will be larger than our overall uncertainty budget. For the NS233 detector, we observe a ≈3.5 % lower system DE than when splicing the fibers. In the extreme case, an FC/PC connection may have very low losses (close to 0 %). Therefore, we speculate that this measurement already reveals a variation of at least 3.5 % in the extracted DE for the FC/PC connector method.

References

[1] T. Gerrits et al., "Calibration of free-space and fiber-coupled single-photon detectors" Metrologia **57**, 015002 (2020)

[2] https://consensus.nist.gov/

A DEEP NEURAL NETWORK MODEL FOR LEARNING RUNTIME FREQUENCY RESPONSE FUNCTION USING SENSOR MEASUREMENTS

Yongzhi Qu[1], Gregory W. Vogl[2], Zechao Wang[3]

[1]University of Minnesota Duluth, MN, US
[2]National Institute of Standards and Technology, Gaithersburg, MD, US
[3]Wuhan University of Technology, Wuhan, China

ABSTRACT

The frequency response function (FRF), defined as the ratio between the Fourier transform of the time-domain output and the Fourier transform of the time-domain input, is a common tool to analyze the relationships between inputs and outputs of a mechanical system. Learning the FRF for mechanical systems can facilitate system identification, condition-based health monitoring, and improve performance metrics, by providing an input-output model that describes the system dynamics. Existing FRF identification assumes there is a one-to-one mapping between each input frequency component and output frequency component. However, during dynamic operations, the FRF can present complex dependencies with frequency cross-correlations due to modulation effects, nonlinearities, and mechanical noise. Furthermore, existing FRFs assume linearity between input-output spectrums with varying mechanical loads, while in practice FRFs can depend on the operating conditions and show high nonlinearities. Outputs of existing neural networks are typically low-dimensional labels rather than real-time high-dimensional measurements. This paper proposes a vector regression method based on deep neural networks for the learning of runtime FRFs from measurement data under different operating conditions. More specifically, a neural network based on an encoder-decoder with a symmetric compression structure is proposed. The deep encoder-decoder network features simultaneous learning of the regression relationship between input and output embeddings, as well as a discriminative model for output spectrum classification under different operating conditions. The learning model is validated using experimental data from a high-pressure hydraulic test rig. The results show that the proposed model can learn the FRF between sensor measurements under different operating conditions with high accuracy and denoising capability. The learned FRF model provides an estimation for sensor measurements when a physical sensor is not feasible and can be used for operating condition recognition.

Keywords: Frequency response function, Encoder-decoder, Neural network, Deep learning

1. INTRODUCTION

Frequency response functions (FRFs), defined as the ratio between input and output spectra, describes the steady-state relationship between each possible sinusoidal input and the corresponding output under zero initial conditions for linear time-invariant (LTI) systems. Learning FRFs between system inputs and outputs is important for system identification and control. An accurate FRF can help with condition-based maintenance and system response prediction of mechanical systems as it models the system dynamics and can be used to monitor system parameters. FRFs are often estimated analytically or obtained experimentally. However, in certain dynamic operations, the FRF can be difficult to predict or obtain due to experimental difficulties, unsteady operating conditions, and mechanical noise. For example, the FRF is often measured between the dynamic force at the tool tip and the dynamic response measured on a stationary point on milling machine tools, such as the spindle headstock. Since the cutting tool is rotating, the transmission path between the tool tip and the spindle headstock involves a flexible component (the cutting tool) and a moving contact surface. The source excitation spectrum from the tool tip can be modulated by bearing ball pass frequencies and low-pass filtered by the flexible component along the transmission path. One frequency component can affect the harmonic frequency components at another measurement point due to mechanical modulation. Also, the structural resonance frequencies can serve as carrier frequencies that shift the original excitation frequency components to a new

Qu, Yongzhi; Vogl, Gregory W.; Wang, Zechao. "A Deep Neural Network Model for Learning Runtime Frequency Response Function Using Sensor Measurements." Paper presented at ASME 2021 16th International Manufacturing Science and Engineering Conference (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021.

frequency. Therefore, in practical systems, especially with relative movement involved between input and output points, the frequency components can affect one another.

Because the FRFs may be difficult to determine for complex systems with moving parts and varying operating conditions, this paper proposes to learn a ***runtime FRF*** from sensor measurements that would model the previously mentioned factors with a data-driven black box model. Such a method would have applications for manufacturing systems in which sensor-based systems are needed to decrease machine downtime and increase product quality and enrich the knowledge of complex manufacturing processes. For example, it is typically not possible to monitor real-time cutting forces during milling processes within machine tools. However, as a future application, by using runtime FRFs to relate force signals to on-machine sensor signals with known cutting force before cutting, on-machine sensors could be set up to monitor cutting forces during real-time machining for quality control purposes.

Note that the *generalized FRF* has been used to define FRFs for nonlinear systems under nonzero initial conditions, especially for nonlinear Volterra systems [2][3]. Generalized FRFs include the effects on the output response from previous inputs. The term of ***runtime FRF*** in this paper is ***different from*** existing definitions for *generalized FRF*. As argued in the above section, this work assumes that in general mechanical systems, one input frequency component may be related and can be used to calculate a different frequency component in the output spectrum. In the generalized runtime FRF, these cross-relationships among different frequencies can be learned.

It is generally understood that one wave frequency cannot affect another wave frequency. Hence, the system response at one frequency can only be determined by the same frequency component in the input spectrum. However, for certain mechanical and electrical systems, there are certain conditions needed for this constraint to hold in practical FRF measurements: 1.) The signals are not modulated internally between inputs and outputs; 2.) Sensors for data collection have wide-enough frequency range to capture the significant spectral amplitudes; and 3.) Nonlinearities in the system are insignificant and do not add extra frequency components to the output response. In practical systems, especially for assemblies containing multiple components, the above assumptions may not hold.

To account for all the previously mentioned factors, we propose to learn the ***generalized runtime FRF*** from sensor measurements with a data-driven black box model. The rest of the paper is organized as follows: Section 2 reviews related work for FRFs and vector regression; Section 3 introduces the basic background of encoder-decoder neural networks and presents the methodology; Section 4 provides the experimental setup for the case study; Section 5 presents the results and data analysis; and Section 6 concludes the paper.

2. RELATED WORKS

Existing research on FRF identification can be separated into several categories: physics-based FRF evaluation, experimental data-based FRF estimation, and hybrid methods for FRF identification. Theoretically, FRFs can be obtained analytically from the governing differential equations of the system [1]. Towards this direction, the effects of different factors on the FRF have been studied. For machine tools, the single-point FRF at the tool tip has been widely studied for the identification of chatter-free zones. The effects of rotational speed [5] and nonlinear behavior [6] have been considered for FRF identification. The effect of rotor rub-impact has also been modeled for the analysis of the FRF of rotor systems [7]. For nonlinear systems with fading memories, Volterra-based FRF models can be established when the system parameters are known using Volterra series [4]. However, Volterra models require the complete knowledge of the system parameters for the governing differential equations [3]. In applicable cases, system parameters can be measured offline or calculated using geometric and material parameters. However, for practical mechanical systems, the runtime boundary conditions and system parameters may change due to external loads and dynamic interactions between moving parts. Therefore, accurate estimation of system parameters are often infeasible.

With regards to experimental methods, impact hammer testing and sinusoidal function sweeping-based FRF identification are available in most commercial modal testing systems. Receptance coupling provides a hybrid approach to identify the FRF for assemblies. Several studies have tried to obtain FRFs for coupled mechanical structures with multiple components. For example, Schmitz and Duncan predicted the output frequency response of an assembly of nested components with common neutral axes using the receptance coupling approach [8].

There have been a couple of attempts towards runtime FRF determination. Kushnir proposed to estimate the runtime FRF for a lathe using the measured vibration spectrum at the top plate and spindle headstock [9]. The author took the average of the input and output spectrums under different spindle speeds, and then took the ratio between the averages to get the dynamic FRF. This ratio assumes no interaction among different frequency components in the input and output spectrums. Thenozhi and Tang used a radial basis neural network and support vector machine (SVM) to learn the frequency response function from simulation data [3]. The input space they used for the learning process was set as the system input amplitude and frequency, and learning outputs were set as the amplitudes of frequency response. They built a regression model between the learning inputs and outputs. While their input space is in the space of R^2 (i.e., two dimensions) and output space in R (i.e., one dimension), their learning model learned the output frequency response as OFR = f(a, w), where a is the amplitude and w is the frequency, and OFR is the amplitude of the corresponding output frequency response. Similar to the model in [9], the model in [3] is equivalent to finding the average ratio between input and output spectrums. Once again, this ratio assumes no interaction among different frequency components and one common average model is obtained for all operating conditions.

Therefore, all existing approaches assume that the FRF function is strictly restricted to one input frequency component affecting only the same frequency component in the output response. In this manner, the FRF is linear and cannot include nonlinearities under different operating conditions. The approach proposed in this paper is dramatically different from previous work, as this paper proposes a vector regression algorithm using neural networks in the whole frequency range, which allows the cross-relationships between different frequencies to be established. Since the model is nonlinear and includes interactions between different frequencies, it allows different ratios between input spectrums and output spectrums at different operating conditions. In other words, the model provides a regression relationship between the vector space of input spectrum and output spectrum. Also, this paper uses practical sensor measurements from mechanical systems with potentially low signal-to-noise ratios. It can provide a robust approach for real-time classification of operating conditions using a real-time input spectrum. As a general vector regression-based FRF identification approach, it can be applied to any system input and output spectrum measurements.

Next, related works in deep learning for vector regression will be reviewed. An encoder-decoder neural network is a special type of neural network that features dimension reduction during the encoding process and signal reconstruction during the decoding process. The most common application of encoder-decoder networks is an autoencoder, where the training input and training output are identical. An autoencoder is often used for feature extraction as an unsupervised algorithm for dimension reduction purposes and has seen many applications [10]. Supervised and semi-supervised autoencoders have also drawn much research attention for concurrent feature extraction and classification tasks [11]. Lei *et al*. proposed that the autoencoder can be used for regression purposes if the input and output are set with different time series as regularizers [12]. Inspired by Lei's work, this paper reveals an encoder-decoder structure for multivariate vector regression. Lei *et al.* used a supervised autoencoder for the classification task, and the encoding-and-decoding network serves as a regularizer. In this paper, we consider the vectors fed to the outputs of the decoder networks as underlying labels in vector form in the target space of regression. Therefore, the encoder-decoder network will simultaneously learn the mapping relationship as well as the regression tasks for different input-output groups.

An encoder-decoder structure has also been applied to sequence-to-sequence learning for speech recognition [13]. However, the structure used for sequence-to-sequence recognition took a recurrent neural network/long short-term memory (RNN/LSTM) structure to consider time-dependent relationships with a sequential memory mechanism other than an autoencoder-like structure, which does not focus on temporal dependency. While our model shares a similar common ground with sequence-to-sequence models, the proposed approach distinguishes itself from the RNN/LSTM models. The proposed encoder-decoder network aims to learn the intra-class and inter-class spatial relationship between R^N to R^N in vector space, while sequence-to-sequence models aim to model the dependency in temporal space inside the sequence vector and then perform regression tasks in the spatial space. Our model assumes a general regression relationship between two vectors without explicitly considering temporal dependency and therefore the application scenarios are different.

3. BACKGROUND AND PROPOSED METHODS

As illustrated in the Introduction section, we formulated the identification of the generalized runtime FRF as a vector regression problem. The regression can be understood in two levels. The first level is that each frequency component in the output spectrum will be determined by a regression model that takes a vector of the whole input spectrum. The second level is that the regression relationship is learned in the vector space for inputs-outputs pairs under different operating conditions, which can be understood as regression in Hilbert space or function space. Such a regression problem can be solved by a deep encoder-decoder neural network. Next, we introduce encoder-decoder networks and propose a customized model for FRF learning.

3.1 Background Introduction and Proposed Encoder-Decoder Network

An encoder-decoder is a type of neural network that first encodes the inputs to a low dimensional space, and then decodes the data back to high dimensional space. One example of encoder-decoder network is an autoencoder, where the inputs and outputs are the same, and it works as an unsupervised dimension reduction model. A more general form of an encoder-decoder network can take different data as inputs and outputs, and serve as a vector regression model.

In this paper, a partially-tied-weight deep encoder-decoder network for vector regression between the input and output spectrums is proposed. The structure of the deep network is shown in Figure 1. In the proposed structure, tied weights are set for the embedded space learning for both the input spectrum and output spectrum, and an independent output embedding layer is added to allow different coordinates in the embedding space for the inputs and outputs, as shown in the red box in Figure 2. Then, the input and output embedding layers are fully connected to learn the mapping relationship between the input and output embeddings. The number of nodes in each layer are: 200, 100, 50, 50, 100, and 200 with a symmetric layout. Layer 1 and Layer 2 share the same weights with Layer 5 and Layer 4, respectively.

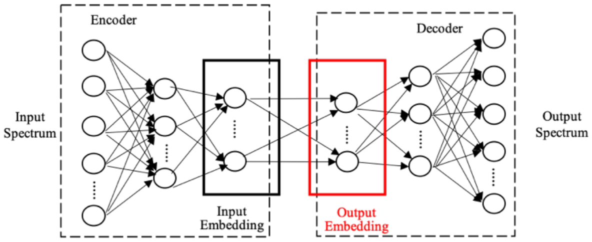


FIGURE 2. PROPOSED ENCODER-DECODER STRUCTURE FOR FRF LEARNING.

The whole network can be represented by the following equations:

$$X_{emb} = f_1(X) \tag{2}$$

$$Y_{emb} = f_2(X_{emb}) \tag{3}$$

$$Y = f_1^{-1}(Y_{emb}), or\ Y_{emb} = f_1(Y) \tag{4}$$

where X is the input of the neural network, Y is the output of the neural network, X_{emb} is the embedded coordinates for X, and Y_{emb} is the embedded coordinates for Y. The proposed encoder-decoder network includes a tied symmetric layer for the learning of input and output embeddings, as well as a fully connected layer from the input embedding to the output embedding, to enable learning of the mapping relationship between the two spectrums.

4. EXPERIMENTAL SETUP

4.1 Test Rig and Sensor Deployments

To validate the proposed FRF learning model, experiments were conducted on a high-pressure hydraulic pipeline system, as shown in Figure 3. For the given hydraulic system, the excitation force comes from the distributed dynamic pressure of the flowing liquid imposed on the inner surface of the pipeline. Measurement of such a distributed dynamic pressure would be physically infeasible. Therefore, in order to measure the excitation force on the pipeline, the dynamic strains on the outer surface of the pipeline were measured with fiber Bragg grating (FBG) sensors. The dynamic strains measured by FBG sensors serve as an indirect measurement of the dynamic pressure inside the pipeline. Using strain sensors for static pressure measurements is a common practice. Dynamic pressure measurements with strain gauges has been recently reported [14]. In this study, since the interested frequency range is about 0 Hz to 1000 Hz, which are the limits of the data collection capabilities, it can be assumed that the strain is linearly related to the dynamic pressure, which is the excitation to the pipeline. Therefore, the strain spectrum was considered to be the input spectrum of the hydraulic system and the FRF learning model, while the vibration spectrum was considered as the output spectrum.

In order to validate the proposed method for learning a generalized runtime FRF under different operating conditions, three sets of experiments were conducted under different operating pressures, namely, 4.62 MPa, 5.65 MPa, or 9.3 MPa. Vibration data and strain data over a section of straight pipeline were collected. A total of about 40 minutes of data were collected. Vibration data were collected using a tri-axial accelerometer and strain data were collected using FBG sensors. There are a total of five sensor measurements: X, Y, and Z accelerations, circumferential strain, and axial strain. For the vibration coordinates, the X axis points down to the ground, the Y axis points axially, and the Z axis points transversely, as shown in Figure 4. The sampling rate for vibration signals is 10240 Hz, and the sampling rate for FBG strains is 2000 Hz. While the sampling frequency for the FBG sensors (limited by the data acquisition system) falls a little short to cover 0 Hz to 1000 Hz, it would not significantly affect our validation process, due to the lower spectral amplitudes near 1000 Hz.

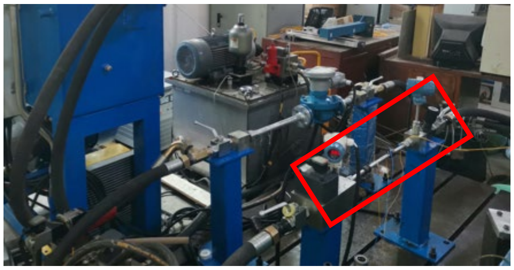


FIGURE 3. EXPERIMENTAL HYDRAULIC SYSTEM.

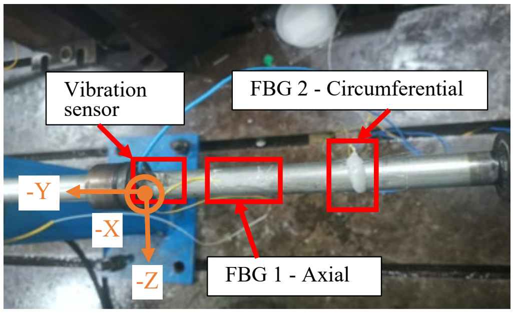


FIGURE 4. SENSOR LOCATIONS.

Based on experimental research under different pressures and quantitative calibration, it was found that the circumferential FBG is more sensitive to the pressure of the pipeline and presents a nearly linear relationship to the hydraulic pressure. Therefore, the FBG 2 sensor was selected for the FRF learning. For vibration, the Z-axis acceleration, which represents the transverse vibration in the horizonal plane and perpendicular to the axial direction of the pipeline, is chosen as the output corresponding to the FBG input.

4.2 Data Preparation

In order to provide rich enough data for FRF learning, each signal collected over about 40 minutes was segmented into multiple samples and then the Fast Fourier Transform (FFT) was used to yield the frequency spectrum of each segment. Since our vibration data and strain data were collected separately with different hardware and software systems asynchronously (misaligned by 10 s of seconds to a few minutes), in order to overcome the synchronization error and obtain more spectrums for training, the original data under the same operating condition was segmented with 50 percent overlap, which means adjacent data segments overlap by 50 percent with each other. Each data segment contains signals collected over 0.2 s, corresponding to 2024 data points for acceleration data and 400 data points for the FBG strain data, in the time domain. Both FFTs for acceleration and strain have a frequency resolution of 5 Hz, which provides 201 spectral data points for the frequency range of 0 Hz to

Qu, Yongzhi; Vogl, Gregory W.; Wang, Zechao. "A Deep Neural Network Model for Learning Runtime Frequency Response Function Using Sensor Measurements." Paper presented at ASME 2021 16th International Manufacturing Science and Engineering Conference (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021.

1000 Hz. The first FFT point, which is for 0 Hz, is removed, which yields 200 frequencies for learning purposes. A total of 23 254 pairs of sample segments for the training and test is obtained, where a strain spectrum and an acceleration spectrum form one pair of sample segments. The whole data sets are summarized in Table 1, in which each sample is a sample segment, a group of sequential data points, collected over 0.2 s.

TABLE 1. NUMBER OF SAMPLES FOR FRF LEARNING.

No. of samples	4.62 MPa	5.65 MPa	9.3 MPa	Total No.
FBG Strain	5242	12 011	6001	23 254
Vibration	5242	12 011	6001	23 254

During the model learning and testing process, about 85 percent of data was used for training and 15 percent was used for testing. Since the samples include overlap, which may help the testing processing if a similar sample with overlap has been seen in the training datasets, a random pool of datasets was not selected for training. Instead, for each of the operating conditions in Table 1, the first 85 percent of the data collected over time was used as training data and the latter 15 percent of data was used for testing. For example, for 4.62 MPa, the first 4500 samples out of the total 5242 samples were utilized for training. The rest of the samples served as testing datasets. A total of 20 000 samples from three different operating conditions were used for training and 3254 samples were used for testing.

4.3 Model Setting

This section provides the setting of our models for the deep encoder-decoder network. Since there exists a hybrid setting for the weight parameter, the total number of parameters for the network is 22 840. This is compared to 44 840 parameters if an un-tied encoder-decoder would have been used. Therefore, the partially-tied weights significantly reduced the number of parameters in the model. A batch optimization algorithm was adopted with a batch size of 100 samples, and the learning rate was set as 0.008. Twenty (20) epochs were performed to have the training process converge when the training loss and validation loss stopped decreasing. A quick overfit check with a 20% validation dataset over 50 epochs indicated that more epochs do not introduce overfitting since the accuracy on the validation set did not change, while the validation loss stayed relatively low after about 20 epochs.

The learning process was conducted on a MacBook Air computer with a 1.8 GHz Intel i5 processor and 8 GB of memory. On average, learning took about 11.8 s for 20 epochs, which is about 0.59 s for each epoch. Therefore, the training process is quite efficient.

5. RESULTS AND DISCUSSION

5.1 Results with the Proposed Model

In this section, the entire learning process is visualized, starting with the raw data and ending with the predicted output response from the learned FRF model. Figure 5 shows plots of the raw acceleration and FBG strain signals, and Figure 6 shows examples of FFTs of sample segments used for model learning. Note that many of the amplitudes in the strain spectrum are about 25% of the peak amplitude shown around 250 Hz. In contrast, the noise floor in the vibration spectrum is about 1% or less of its peak amplitude around 250 Hz; the vibration sensor has a much higher signal-to-noise ratio. Also, there is one dominant frequency component, which is suspected to be the first natural frequency of the pipeline structure that should capture the majority of vibrational displacements. The strain spectrum contains more noise and is flatter compared with the vibration spectrum. Yet, both spectrums share the same dominant frequency.

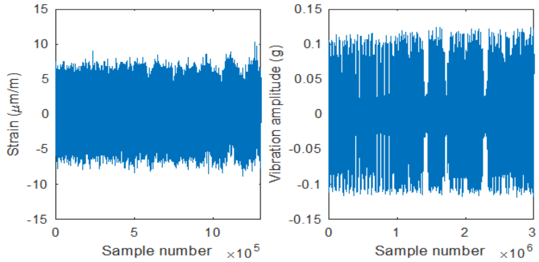


FIGURE 5. SAMPLE STRAIN AND VIBRATION SIGNALS.

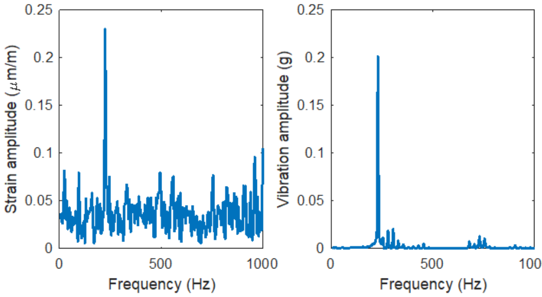


FIGURE 6. EXAMPLE FFTS OF STRAIN AND VIBRATION SIGNALS.

Without further processing of the data before use in the neural network, the peak frequency components would dominate over all other frequency components with regards to training errors. During the training process to map strain (the input) to vibration (the output), the low-amplitude spectral components would be neglected, since the error contributions from smaller amplitudes are much lower than those for higher amplitudes. Thus, the natural log of vibration FFT amplitude is set as the output for the neural network, which enables network learning to be influenced by all frequency content. A similar approach has been used in another method to allow all frequency content to influence the FRF-based solution [15]. Figure 7 shows the same sample spectrum as in Figure 6, except that the vibration spectrum is the natural log of its original spectrum. As one can see from Figure 7, the vibration spectrum is flattened by the log operation and shows more details for the lower-amplitude frequency components. The spectrums are then normalized to [0, 1] to eliminate scalings and were fed into the deep neural

Qu, Yongzhi; Vogl, Gregory W.; Wang, Zechao. "A Deep Neural Network Model for Learning Runtime Frequency Response Function Using Sensor Measurements." Paper presented at ASME 2021 16th International Manufacturing Science and Engineering Conference (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021.

network model. The normalization process rescales the data to a common range. Also, since Sigmoid is used as the activation function in this work, output of the neural networks needs to be normalized to [0,1].

Next, the learning results with the proposed encoder-decoder regression model are discussed. Figure 8 shows an example of the predicted outputted acceleration spectrum tested with the inputted strain spectrum for a sample segment. It can be seen that the predicted vibration spectrum closely follows the actual vibration spectrum corresponding to the testing data. However, the actual spectrum has more fluctuations, while the predicted spectrum is smoother. While it can be expected that the predicted spectrum presents an 'averaging' effect, it is surprising to find that output spectrums calculated for all test samples under the same operating condition are almost identical. This indicates that the trained deep encoder-decoder networks have excellent denoising capability. We are further interested in evaluating whether the predicted spectrums correctly reflect and separate with different operating conditions, which can help us evaluate the generalization capability of the trained deep neural network.

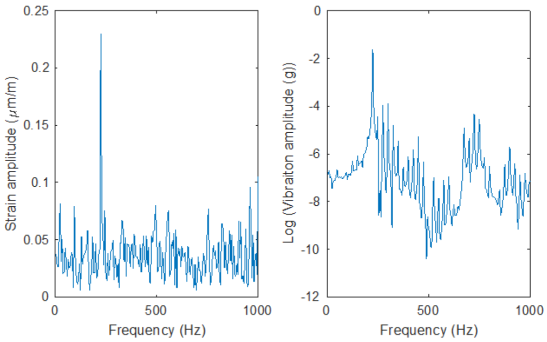


FIGURE 7. STRAIN SPECTRUM AND LOG-OF-VIBRATION SPECTRUM.

Since there are three different operating conditions, it is shown next whether the predicted output response correctly reflects the changing operating conditions, which represent the regression performance in the vector space. Figure 9 shows the predicted spectrums for three different operating conditions, a.k.a. 4.62 MPa, 5.65 MPa, and 9.3 MPa. Again, extensive examinations reveal that all predicted spectrums for one particular operating condition are almost identical even if different sample segments are used as the testing input. However, as Figure 9 shows, the learned generalized FRF captured the real difference between different operating conditions. Figure 9 shows that the predicted spectrum of 9.3 MPa clearly separates itself from those for 4.62 MPa and 5.65 MPa. The results for 4.62 MPa and 5.65 MPa can hardly be separated from each other. The reason is explained in that the data for 4.62 MPa and 5.65 MPa were collected from the same hydraulic system with a setting of 5 MPa. On the other hand, for 9.3 MPa, the control parameter was actually set to 10 MPa.

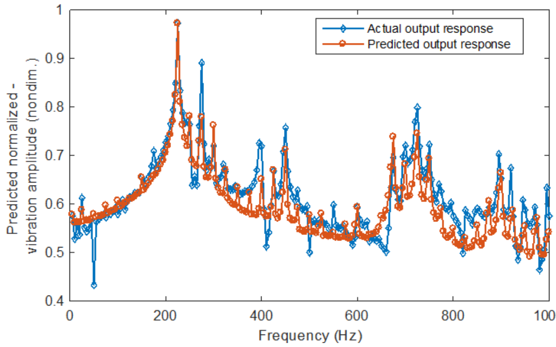


FIGURE 8. EXAMPLE OUTPUT SAMPLE SPECTRUM AND ITS PREDICTED SPECTRUM.

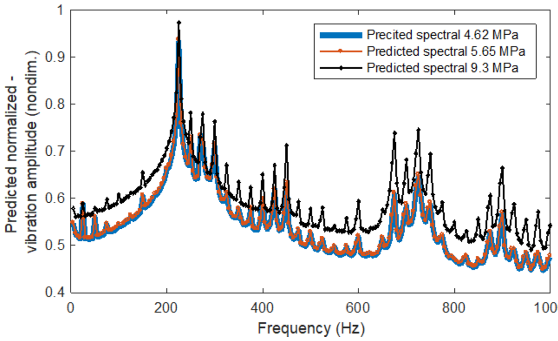


FIGURE 9. PREDICTED VIBRATION SPECTRUMS FOR THREE OPERATING CONDITIONS.

To further validate the results, the averaged spectrum for each of the three operating conditions are shown in Figure 10. The averaged spectrum is the average over all training and testing data. It is perhaps surprising to see that the learned spectrums approximate the average spectrums. Note that the amplitude of the averaged spectrum is slightly lowered due to the averaging effect over non-strictly aligned data, which is typical for averaging of FRFs with noise. Once again, the actual spectrums of 4.62 MPa and 5.65 MPa are very close to each other. Considering that the strain FRFs (the inputs) are extremely noisy with minor differences in the training targets fed to the model, the model cannot learn discriminative outputs for 4.62 MPa and 5.65 MPa. To better explain the averaging and denoising effects of the proposed model for FRF learning, examples of testing input spectrums and the corresponding output spectrums are plotted in Figure 11.

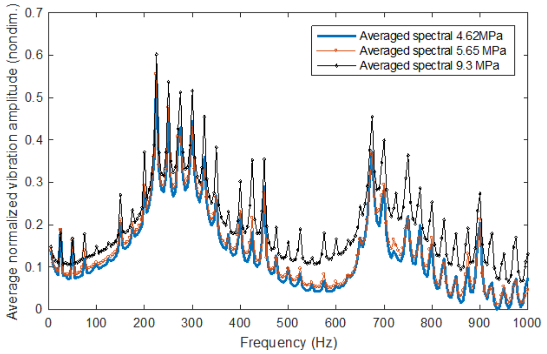


FIGURE 10. AVERAGE SPECTRUM FOR EACH OF THE THREE SEPARATE OPERATING CONDITIONS.

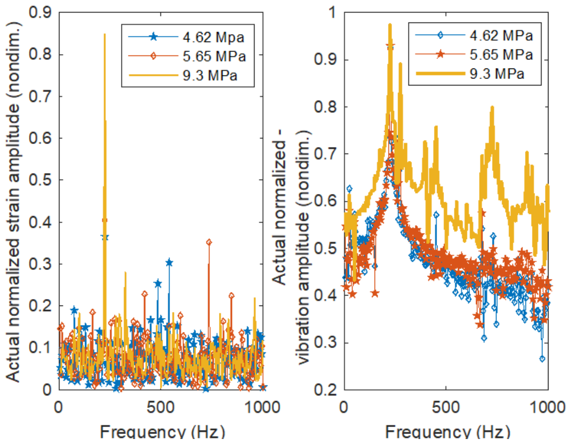


FIGURE 11. SPECTRUMS FOR AN ACTUAL SAMPLE USED FOR TESTING.

It can be seen from Figure 11 that the input spectrums are extremely noisy and the actual corresponding vibration spectrums are relatively noisy, as well. Furthermore, the vibration FFT amplitudes for 4.62 MPa are slightly lower than those for 5.65 MPa, and both are clearly separated from the FFT for 9.3 MPa. Note the differences of the average predicted spectrums in Figure 10 with the actual vibration spectrums in Figure 11; the noise-related fluctuations are filtered during the learning process, which justify the robustness of the proposed deep encoder-decoder network.

5.2 Comparison with Shallow Neural Networks

Next, the accuracy of the generalized runtime FRF is quantified. For various samples, the intra-group predictions under the same operating condition should be very close to each other, since they are sampled from the same process. In contrast, the inter-group predictions should be distinguishable in order to separate responses due to different operating conditions. Therefore, it is expected that the learned model could correctly predict discriminative results under different operating conditions.

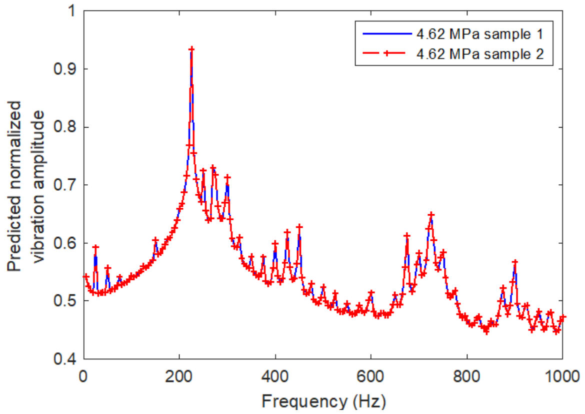


FIGURE 12. RESULTS COMPARISON FOR DIFFERENT TEST SAMPLES UNDER THE SAME OPERATING CONDITIONS FOR THE PROPOSED DEEP MODEL.

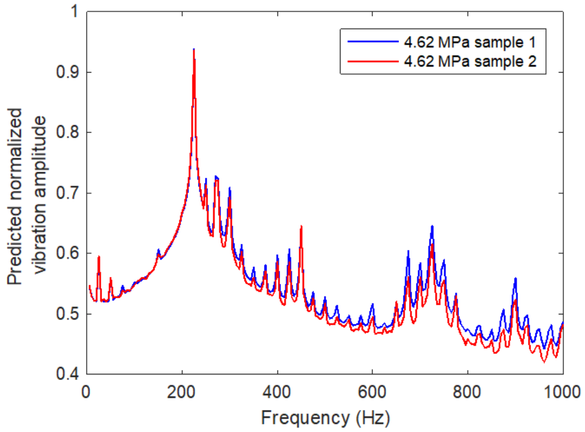


FIGURE 13. RESULTS COMPARISON FOR DIFFERENT TEST SAMPLES UNDER THE SAME OPERATING CONDITIONS FOR THE SHALLOW NEURAL NETWORK.

Figure 12 and 13 present the prediction results for the same two samples from the same operating conditions using the proposed deep neural network and a shallow neural network, respectively. It can be seen from Figure 12 and Figure 13 that the deep neural network performs much more robustly to produce almost identical results for the same process with noise filtering capability. However, the results from the shallow neural network present more fluctuations even if they come from the same process. For the prediction results of 9.3 MPa, the results from both the proposed deep model and the shallow network show fluctuations. However, the deep model still presents more robustness.

Also, Figure 9 shows that the deep neural network can predict clearly separable results for different operating processes. In order to quantify the performance of the proposed deep neural network and shallow neural network, we calculated and summarized the in-class fluctuations in terms of standard deviation. Since the standard deviation (STD) is calculated over multiple vectors, the largest STD over all dimensions is used to indicate the fluctuation, as shown in Table 2. Similar results can be obtained for the average standard deviation as well.

TABLE 2. ROBUSTNESS OF PREDICTION INDICATED BY IN-CLASS STD.

In-class STD	Class – 1 (4.62 MPa)	Class – 2 (5.65 MPa)	Class - 3 (9.3 MPa)
Proposed deep NN	0.0083	0.0069	0.0238
Shallow NN	0.0200	0.0145	0.0310

Next, we show the separation capability of predicted results from the proposed deep and shallow neural networks. The estimated FRFs using trained neural networks serve as underlying labels to each operating condition. However, since there are no true labels for them, a k-mean clustering task is performed over both networks to see whether the predicted results are discriminative with regards to the operating process or not. Therefore, we can evaluate whether the predicted results can detect a pressure change in the pipeline.

The k-mean clustering results are shown for all conditions in Table 3, but because Class – 3 corresponds to the 9.3 MPa operating condition and has more noise in the test data, Table 4 shows the separation accuracy for only Class – 3. It can be seen in Table 3 and Table 4 that the proposed deep regression model based on encoder-decoder networks clearly outperforms the shallow neural network with both FRF prediction robustness and discriminative capability.

TABLE 3. GROUPING ACCURACY FROM CLUSTERING RESULTS (ALL).

Comparison of FRF prediction accuracy, All	Deep Encoder-decoder NN	Shallow NN
k-means clustering accuracy	98.59%	97.79%

TABLE 4. GROUPING ACCURACY FROM CLUSTERING RESULTS (9.3 MPA).

Comparison of FRF prediction accuracy, Class - 3	Deep Encoder-decoder NN	Shallow NN
k-means clustering accuracy	95.51%	92.91%

6. CONCLUSION

In this paper, identification of a generalized runtime FRF under multiple operating conditions was proposed as a supervised vector regression problem. The proposed model was tested for regression performance in the vector space to classify different inputs under changed operating conditions. A deep encoder-decoder-based model was proposed for the learning task. The model features partially-tied weights for the encoding and decoding process. An added layer for the output embedding was proposed to store different coordinates under the embedded space. The output embedding is fully connected with the input embeddings, which creates the mapping/regression relationship between them in the embedded space.

The model was validated using experimental measurement data from a hydraulic system, in which the relationship to be learned is between the vibration response (measured by an accelerometer) and the pressure (measured by an FBG strain sensor). The results showed that the proposed model can effectively learn the FRF function between the input and output spectrums under different operating conditions. The learned model also demonstrated excellent denoising capability, which was not shown by other FRF approaches. The presented work was for learning and classification of a system under very controlled conditions, and as such, is the first step of a longer process that will result in the generalized runtime FRF for prediction purposes.

NIST DISCLAIMER

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States. Approved for public release; distribution is unlimited.

ACKNOWLEDGEMENTS

Yongzhi Qu acknowledges the Department of Commerce of the United States for partial support of this research under contract No. 70NANB20H175, and partial support from the University of Minnesota in Grant-in-Aid of Research, Artistry and Scholarship (GIA).

REFERENCES

[1] Liu MY, Wang ZC, Zhou ZD, et al. Vibration response of multi-span fluid-conveying pipe with multiple accessories under complex boundary conditions. Eur J Mech A-Solid 2018;72:41-56.

[2] Lang, Z.-Q., & Billings, S. (1996). Output frequency characteristics of nonlinear systems. International Journal of Control, 64(6), 1049–1067.

[3] S. Thenozhi, Y. Tang, Learning-based frequency response function estimation for nonlinear systems, Int. J. Syst. Sci. 49 (2018) 2287–2297.

[4] Xing Jian Jing, Zi Qiang Lang, Stephen A. Billings, Output frequency response function-based analysis for nonlinear Volterra systems, Mechanical Systems and Signal Processing, Volume 22, Issue 1, 2008.

[5] Niccolò Grossi, Lorenzo Sallese, Filippo Montevecchi, Antonio Scippa, Gianni Campatelli, Speed-varying Machine Tool Dynamics Identification Through Chatter Detection and Receptance Coupling, Procedia CIRP, Volume 55, 2016, pp. 77-82.

[6] A.S. Delgado, E. Ozturk, N. Sims, Analysis of Non-linear Machine Tool Dynamic Behavior, Procedia Engineering, Volume 63, 2013, pp. 761-770.

[7] Y. Liu, Y.L. Zhao, J.T. Li, H. Ma, Q. Yang, X.X. Yan, Application of weighted contribution rate of nonlinear output frequency response functions to rotor rub-impact, Mechanical Systems and Signal Processing, Volume 136, 2020.

[8] Tony L. Schmitz, G. Scott Duncan, Receptance coupling for dynamics prediction of assemblies with coincident neutral axes, Journal of Sound and Vibration, Volume 289, Issues 4–5, 2006, Pages 1045-1065.

[9] Kushnir, E. "Determination of Machine Tool Frequency Response Function During Cutting." Proceedings of the ASME 2004 International Mechanical Engineering Congress and Exposition. Pressure Vessels and Piping. Anaheim, California, USA. November 13–19, 2004. pp. 63-69. ASME.

[10] Qu, Y.; He, M.; Deutsch, J.; He, D. Detection of Pitting in Gears Using a Deep Sparse Autoencoder. Appl. Sci. 2017, 7, 515.

[11] Ranzato, M. and Szummer, M. (2008). Semi-supervised learning of compact document representations with deep networks. In Proceedings of the 25th International Conference on Machine Learning (ICML), pages 792–799.

[12] Lei Le, Andrew Patterson and Martha White, Supervised autoencoders: Improving generalization performance with unsupervised regularizers, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

[13] Ilya Sutskever, Oriol Vinyals, Quoc V. Le, Sequence to Sequence Learning with Neural Networks, NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, Cambridge, MA, USA.

[14] Fatima Garcia Castro, Olivier de Sagazan, Nathalie Coulon, Antoni Homs Corbera, Dario Fassini, Jeremy Cramer, France Le Bihan, μ-Si strain gauge array on flexible substrate for dynamic pressure measurement, Sensors and Actuators A: Physical, Volume 315, 2020.

[15] Gregory W. Vogl, M. Alkan Donmez, A defect-driven diagnostic method for machine tool spindles, CIRP Annals, Volume 64, Issue 1, 2015, pp. 377-380.

Qu, Yongzhi; Vogl, Gregory W.; Wang, Zechao. "A Deep Neural Network Model for Learning Runtime Frequency Response Function Using Sensor Measurements." Paper presented at ASME 2021 16th International Manufacturing Science and Engineering Conference (MSEC 2021), Cincinnati, OH, US. June 21, 2021 - June 25, 2021.